USING MULTIDIMENSIONAL ITEM RESPONSE THEORY TO REPORT SUBSCORES ACROSS MULTIPLE TEST FORMS

By

Jing-Ru Xu

A DISSERTATION

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

Measurement and Quantitative Methods—Doctor of Philosophy

2016

ABSTRACT

USING MULTIDIMENSIONAL ITEM RESPONSE THEORY TO REPORT SUBSCORES ACROSS MULTIPLE TEST FORMS

By

Jing-Ru Xu

There is an increasing interest in subscores in educational testing because subscores have potential benefits in remedial and instructional application (Sinharay, Puhan, & Haberman, 2011). Users of score reports are interested in receiving information on examinees' performances on subsections of an achievement test. These scores "typically are referred to as 'subscale scores,' 'subtest scores,' or more generically, 'subscores (Ferrara & DeMauro, 2006, p. 583).'"

Among these current subscore research reports, few address the following issues. First, in most research, the number of subscores, the number of items in each subscore domain and the item types in each domain are already fixed according to the classification produced by test developers and content experts. Thus, the distinct domains defining subscores may not be clearly defined in a technical psychometric sense. Also, little information may be provided to show there are enough items in each domain to support reporting useful scores. Moreover, it may not be clear why particular types of items are grouped together within each domain. Finally, few discuss how to link and equate test forms when reporting subscores.

In order to fill in the above gaps and to explore solutions to the questions, this research study applied the multidimensional item response theory to report subscores for a large-scale international English language test. Different statistical and psychometric skills and methods were used to analyze the dimension structure, the clusters for reporting subscores, and to link individual test forms to provide comparable and reliable subscores.

The results show that there are seven distinct dimensions that capture the variation among examinee responses to items in the data sets. For each different form, there are different number of clusters identified. Moreover, each cluster corresponds with a unique reference composite. Across all five test forms, there are 6 – 8 clusters identified. There is a consistency of the dimensional structure across these five forms based on the parallel analysis, exploratory and confirmatory factor analysis, cluster analysis and reference composite analysis. The nonorthogonal Procrustes rotation linked each individual form with the base form and rotated the subscores from individual forms back to the same base form so that the subscores identified from different forms were comparable.

In conclusion, this research provided a systematic method to report subscores using multidimensional item response theory. Such procedures can be replicated and applied for different test programs. Large amount of missing values and small sample size for each individual form were limitations in this study. For future research, I would suggest using large-scale data sets with few missing values. For each individual test form, the sample size should be better larger than 450, such as 600 to 800.

ACKNOWLEGEMENTS

It never occurred to me I would be on the final stage of writing the acknowledgements of my dissertation. In fact, being a Ph.D. candidate had never ever been my life goal. The turning point was the first time I heard about item response theory. When Fred Lord first raised this idea, probably no one even himself would expect how it would change the modern measurement and testing theory in society. Human beings' abilities are no longer judged or measured based on the final composite score but determined by the information contained in individual test items.

At the very beginning, I had little research experience nor measurement courses taken in this field. The only thing I had was passion, dream, and curiosity. Before entering the psychometrics field, I had been wandering around years to figure out how to help people understand themselves and to measure those invisible abilities – mental, emotional, psychological and any other interesting but hidden powers deep inside human brains.

It was not until I met my advisor Dr. Mark Reckase and started taking courses from Measurement and Quantitative Methods (MQM) program that the veil was lifted. I would express my deepest gratitude to Dr. Reckase who led me into this splendid world, who has always been there guiding and supporting me, and leading me through all those challenges. Without his support, nothing would have been achieved. Especially, when he led me into the multidimensional IRT world, it was fascinating! My dissertation was just the beginning of this journey, and there are tons of thousands of interesting things out there for me to explore. It is as if Dr. Reckase planted those magic seeds inside my mind and during the past years they have been growing and they are waiting for blossom in future.

Finally, I want to thank my parents, especially my mom for the great support and who always has faith into me. I want to give credits to my friends who share laughs and tears with me. I want to thank my colleagues – Joe, Shu-chuan, John, and Will who share views and perspectives on my dissertation. It is a long journey and I feel blessed to have all of these precious people alongside with me. As always, "Two roads diverged in a wood, and I – I took the one less traveled by, and that has made all the difference."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AERA	American Educational Research Association
APA	American Psychological Association
CFA	Confirmatory Factor Analysis
EFA	Exploratory Factor Analysis
ESL	English as a Second Language
ICC	Item Characteristic Curve
IRT	Item Response Theory
MIRT	Multidimensional Item Response Theory
NCLB	No Child Left Behind
NCME	National Council on Measurement in Education
PTEA	Pearson Test of English Academic

CHAPTER 1
INTRODUCTION OF THE RESEARCH QUESTIONS

It all starts with an interesting idea – how to report valid and reliable subscores. In other words, this research targets the consistency, the reliability, and the interpretability of subcores across different test forms within a multidimensional score space. Subscores are scores for different sub-categories or sub-constructs. There are more than one hypothetical construct to be measured for a test. Meanwhile, for testing security purpose, a test program should have more than one test forms. These different test forms are supposed to measure the same multiple latent abilities, skills, knowledge or constructs from examinees. Therefore, subscores from examinees on different test forms should be interchangeable and comparable. In this research, a new systematic methodology was developed to provide solutions to these questions as well as to share insights on the application of subscore reporting in educational measurement field.

1.1 Popularity of Subscores

There is an increasing interest in subscores in educational testing because subscores have potential benefits in remedial and instructional application (Sinharay, Puhan, & Haberman, 2011). Users of score reports are interested in receiving information on examinees' performances on subsections of an achievement test. These scores "typically are referred to as 'subscale scores,' 'subtest scores,' or more generically, 'subscores (Ferrara & DeMauro, 2006, p. 583).'" For example, instructors and parents are interested to know whether English as a Second Language (ESL) students perform speaking as well as writing in an English language test.

Policy makers, college and university admissions officers, school district administrators, educators, and test takers all want subscores to help them make decisions for both admission and diagnosis purposes (Monaghan, 2006). The National Research Council report "Knowing What Students Know" (2001) emphasizes that the goal of assessment is to provide useful information

for examinees' knowledge, skills and abilities. Also, the U.S. Government's No Child Left Behind (NCLB) Act of 2001 requires that students should receive diagnostic reports that allow teachers to address specific diagnostic needs. Subscores can be used to identify such particular information for examinees and to report diagnostic analyses for teachers (Sinharay et al., 2011).

According to Sinharay et al. (2011), various researchers have proposed different methods for examining whether subscores have adequate psychometric quality. For example, Stone, Ye, Zhu, & Lane (2010), Wainer et al. (2001), and Sinharay, Haberman, and Puhan (2007) applied different factor analysis procedures to explore the distinctiveness of subscores. Harris and Hanson (1991) used the beta-binomial model to analyze whether subscores have added-value over the total score. Another approach to address this issue is to use a multidimensional item response theory (MIRT) model (e.g., Reckase, 1997; Ackerman, Geierl, & Walker, 2003) to analyze the structure of the item response data. See von Davier (2008), Haberman and Sinharay (2010), and Yao et al. (2007) for a detailed description of such methods. Also, Ackerman and Shu (2009) used the dimensionality assessment software programs such as DIMTEST (Stout, 1987) and DETECT (Zhang & Stout, 1999) to identify subscores. Haberman (2008a) and Sinharay (2010) used classical-test-theory-based methods to determine whether subscores have added value over the total score.

Most of the research regarding subscore reporting takes one of two approaches. One focuses on the application of dimensionality analysis using such procedures as factor analysis, MIRT models, and dimensionality assessment software programs to identify subscores for tests that were constructed to yield a well-supported total score (essential unidimensionality). The other approach focuses on the classical-test-theory-based methods, such as those implemented by Haberman and Sinharay's.

1.2 Myth of Reporting Subscores in Current Research

However, among these current subscore research reports, few address the following issues. First, in most research, the number of subscores, the number of items in each subscore domain and the item types in each domain are already fixed according to the classification produced by test developers and content experts. As a result, the distinct domains defining subscores may not be clearly defined in a technical psychometric sense. Also, little information may be provided to show there are enough items in each domain to support reporting useful scores. For example, are 16 items in a particular domain sufficient to represent the skills and knowledge included in that domain? Moreover, it may not be clear why particular types of items are grouped together within each domain. The analyses focus more on supporting the subjective classification of items into domains rather than determining the sets of items that form coherent sets that merit reporting as subscores.

According to the Standard 5.12 of the *Standards of Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), "Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established". Also, the Standard 1.12 further clarifies:

> When a test provides more than one score, the distinctiveness of the separate scores should be demonstrated, and the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed. (p. 20)

The requirements implied by these standards stimulated the following questions. Can the subscore structure defined by the test specifications be supported with empirical evidence about the number of domains, the relationship of item types to domains, and the number of items

within each domain? If the dimensional structure of the test is not appropriately identified, then reporting subscores based on such fixed domains may not be meaningful.

Second, almost no research focuses on the dimensional structure of the test across multiple forms. In most research, the data sets are either from simulations or from a single set of real test data without the number of missing values minimized. Using multiple forms to support the inferences about the dimensional structure of the test for reporting subscores is very important for showing the generalizability of the results. This is especially important when subscores are reported for diagnostic purposes for multiple groups of examinees with differences in demographic and language background.

Third, research articles do not emphasize that the dimensionality needed to model the response data from a test is not only a feature of the test items but also a function of the dimensions of variability of the examinees. The reason for reporting subscores is to diagnose distinctive abilities or skill levels of examinees. Thus, when considering the dimensionality of the response data from a test that is supposed to support subscores, we need to take into account the characteristics of the examinees. Reckase (2009) showed that "the number of dimensions needed to model the item response matrix may be different depending on the amount of variability on the dimensions in different samples of examinees (p.183)."

Furthermore, in most research studies the subscores are considered for one test form. In reality, there are multiple test forms designed and pre-equated for examinees taking on different test dates. Instructors, policy makers, and school administrators want to compare the subscores across different examinee groups in different areas on different test dates. Therefore, linking and equating multiple test forms when reporting subscores is an essential procedure to ensure the validity and reliability of score reporting.

During test development, multiple test forms are often equated under one hypothetical construct, which is based on a unidimensional model. However, when reporting subscores, there are multiple hypothetical constructs. Each test form measures more than one hypothetical construct. The unidimensional assumption is violated. Therefore, how to link and equate test forms using multidimensional item response theory (MIRT) when test forms were originally designed and equated using unidimensional IRT is an interesting part for this research.

Given the issues identified in the research literature on subscores, three major research questions were identified as the focus for the research reported here. First, can MIRT methods be used to identify a reasonable subscore structure for a large-scale test that is well fit by a unidimensional model? This question is addressed in the context of real test data with multiple test forms and the associated problems of missing data. Second, is there evidence that the multidimensional subscore structure generalizes over multiple forms from the same test? Finally, if a subscore structure is identified across multiple test forms, how to link and equate multiple forms to report meaningful subscores that are comparable and interchangeable across different test forms for different examinee groups.

1.3 Research Questions

The data for this research came from a relatively new test of English for those who have other first languages – the *Pearson Test of English Academic (PTEA)* (Pearson Longman, 2010). This test was selected for analysis because it has thorough coverage of the components of English language and item response data were available from individuals from a number of different countries and language backgrounds. There are some complexities in the use of the data from this program, however. The PTEA has many different test forms and a complex

pattern of common items between forms. This makes the analyses of the data from this program challenging. However, through careful analysis, these challenges were overcome and the data were used to address the following specific research questions.

1. How many distinct dimensions are needed to accurately describe the relationships between the test items for the current heterogeneous sample of examinees? In particular, is more than one dimension needed?
2. If more than one dimension is needed to represent the relationships among the test items for the current sample of examinees, are there clusters of items that are sensitive to distinct combinations of skills and knowledge and are these clusters related to known constructs of language performance?
3. If meaningful clusters can be identified, are they specific to one form of the test or do similar clusters appear for more than one form? That is, do multiple forms replicate the complex structure of the language constructs?
4. If replicable clusters can be identified in each test form, how to link and equate different test forms so that subscores from examinees taken in different places on different dates can be comparable and interchangeable?

The results of investigations related to these research questions were used to determine if it is meaningful to report subscores on a large scale test with multiple test forms even though the item response data are well fit by a unidimensional item response theory model when the full examinee sample that is composed of multiple groups is analyzed. Multidimensional item response theory (MIRT) was the main methodology for investigating the research questions.

CHAPTER 2
LITERATURE REVIEW

2.1 Item Response Theory (IRT)

"Item response theory (IRT) is a family of statistical models used to analyze test item data" (Yen and Fitzpatrick, 2006, p. 111). IRT estimates the characteristics of test and examinees using a statistical procedure and states how these characteristics interact in defining item and test performance. IRT models describes the relationship between item scores and examinee ability levels and item parameters using nonlinear functions. The core of IRT models is to relate the probability of getting an item right to an examinees' abilities given the particular responses to an individual item. It is convenient to assume that the responses to individual items are conditionally independent. Lord (1980) states the local independence principle as the probability of success on item *i* given θ is equal to probability of success on item *i* given both θ and the examinee's performance on items *j, k,*…, and so forth. For three items *i, j, k,* the mathematical equivalent for local independence is

$$P\,(u_i = 1, u_j = 1, u_k = 1|\theta) = P(u_i = 1|\theta)\; P(u_j = 1|\theta)\; P(u_k = 1|\theta) \qquad (1)$$

2.1.1 Types of Item Data

Different types of item responses are associated with different item scores. Item scores can be dichotomous (having only two possible outcomes – either correct or incorrect). They can also be polytomous (having more than two possible outcomes). Most constructed-response items or open-ended response items have more than two score categories. For example, polytomous item score rubrics could be 0 = inaccurate answer, 1 = partially correct answer, 2 = completely accurate answer; for rating scales, 1 = completely disagree, 2 = disagree somewhat, … , 5 = strongly agree (Yen and Fitzpatrick, 2006, p. 112).

2.1.2 Dimensionality

IRT models use examinee parameters, such as person parameters, traits, proficiencies, or abilities to describe the dimensions representing important differences in examinees' performances measured by the test items. These dimensions are referred to as "abilities". In educational measurement field, they can be called proficiencies, knowledge, skills, attitudes, or other characteristics. According to Yen and Fitzpatrick (2006), models that use only one ability to quantify the differences among examinees and among items are unidimensional IRT models. Models that use more than one ability are multidimensional IRT models (p. 112).

2.2 Unidimensional IRT (UIRT) Models for Items with Two Score Categories

2.2.1 One-Parameter Logistic Model

The simplest commonly used UIRT model is one-parameter logistic model. It uses one parameter to describe the item characteristics and one parameter to measure person ability. This model can be represented by

$$P\left(U_{ij}=1\middle|\theta_j,b_i\right)=\frac{e^{\theta_j-b_i}}{1+e^{\theta_j-b_i}}, \quad (2)$$

where u_{ij} is the score for Person j on Item i. 1 means the examinee answers the item correct. θ_j is the person characteristics parameter that describes Person j's latent ability or achievement level on a continuous scale related to the performance on Item i. b_i is the item characteristics parameter that describes the item difficulty. The larger the b_i value, the more difficult the item. P is the probability of Person j answering Item i correctly. If person A is more capable than person B, then for the same item with equal b_i value, person A has a higher probability – larger P – than person B.

2.2.2 Two – Parameter Logistic Model

Birnbaum (1968) proposed the two-parameter logistic model to introduce a slightly complex concept of the discrimination parameter, a_i. The mathematical expression for the model is

$$P(U_{ij} = 1|\theta_j,\ a_i, b_i) = \frac{e^{\ a_i(\theta_j - b_i)}}{1+e^{\ a_i(\theta_j - b_i)}}, \tag{3}$$

where a_i is the discrimination parameter and the other symbols have the same definition as those given in one-parameter logistic model. The discrimination parameter reflects the strength of the relationship between probability of correct response and person ability. It indicates the rate of change in probability regarding the unit change of ability scale. A large discrimination parameter means a small change in ability will result in a big change of probability for correct response. It shows how strongly an item can discriminate the ability level when item difficulty does not change.

2.2.3 Three – Parameter Logistic Model

Within person-item interaction there is another feature that indicates low ability examinees can still have the possibility to get an item correct. This characteristic of an item is named as its "guessing" parameter. It represents the empirical observation that a person can get an item right by guessing one of the options from a multiple-choice item. The mathematical formula for the probability of a correct response to the item is given by

$$P(U_{ij} = 1|\theta_j,\ a_i, b_i, c_i\ ) = c_i + (1 - c_i)\frac{e^{\ a_i(\theta_j - b_i)}}{1+e^{\ a_i(\theta_j - b_i)}}, \tag{4}$$

where c_i is the guessing parameter. Therefore, different from the other two UIRT models, the asymptotic probability with extremely low ability examinees to get an item right is no longer

zero but c_i. It is very unlikely that the probability of a person with even no knowledge of the correct answer to get a multiple-choice item right will be zero.

2.3 Unidimensional IRT Models for Items with More than Two Score Categories

Different IRT models are used to describe different item data. Since the required responses for different types of items are different – generating writing samples, solving mathematics problems, and rating statements, the IRT models that describe the item/response interactions are different as well (Reckase, 2009, p. 32). The most commonly used IRT models for polytomous items are – the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded response model (Samejima, 1969). The partial credit model family is the focus of this research.

2.3.1 The Partial Credit Model

The partial credit model is appropriate for items that require successful accomplishment of a number of tasks. It is designed for items with two or more ordered-score categories. The partial credit model considers each score category as correct/incorrect or accomplished/not accomplished. In order to receive the maximum score, examinees need to complete all tasks correct. Therefore, the scores on the item represent different levels of performance. The higher score indicates the examinees have accomplished more desired tasks. The boundaries between adjacent score categories are called thresholds. An examinee's performance is associated with a particular probability on either side of a threshold. At each threshold, the item can be scored dichotomized reflecting the probability of a response either above or below the selected threshold corresponding to a particular score category.

The mathematical formula of the partial credit model is

$$P(u_{ij} = k|\theta_j) = \frac{e^{[\sum_{u=0}^{k}(\theta_j - \delta_{iu})]}}{\sum_{v=0}^{m_i} e^{[\sum_{u=0}^{v}(\theta_j - \delta_{iu})]}}, \quad (5)$$

where k is the score on Item i, m_i is the maximum score on Item i, and δ_{iu} is the threshold parameter for the uth score category for Item i. The threshold parameter shows where the adjacent score categories have equal likelihoods.

2.3.2 The Generalized Partial Credit Model

Muraki (1992) first proposed the generalized partial credit model. It is an extension of partial credit model with the addition of the discrimination parameter, a, to the former one. The difference between the two is similar to the difference between the one-parameter logistic model and two-parameter logistic model. The mathematical formula of the generalized partial credit model is

$$P(u_{ij} = k|\theta_j) = \frac{e^{\left[\sum_{u=1}^{k} Da_i(\theta_j - b_i + d_{iu})\right]}}{\sum_{v=1}^{m_i} e^{[\sum_{u=1}^{k} Da_i(\theta_j - b_i + d_{iu})]}}, \quad (6)$$

where k is the score on Item i, m_i is the maximum score on Item i, and d_{iu} is the threshold parameter for the uth score category for Item i. b_i is the overall difficulty of the test item and a_i is the overall discrimination parameter of the item. It is assumed to be the same across all thresholds but can be different across different items.

2.4 Multidimensional Item Response Theory (MIRT)

MIRT is a model or theory that idealizes the psychological and educational measurement in reality. It approximates the relationship between peoples' capabilities and responses to test items. In other words, it states the relationship between people's locations in a multidimensional

space and the probabilities of their responses to a test item (Reckase, 2009, p. 58). The mathematical models that represent such relationship are MIRT models because they assume multiple hypothetical constructs influence people's performances instead of only one hypothetical construct (Reckase et al., 1988).

If unidimensional IRT is designed to capture the dominant capability of a person, then MIRT is designed to dig deeper to discover the multiple capabilities of a person given the responses from a test item. How the human mind works is a timelessly interesting and fascinating topic for centuries. Tracing back millennia, Plato teased out why human beings could come to know things we had not known before. Aristotle endeavored to virtually encompass all facets of intellectual inquiry. Kant set the foundation of how human mind would structure human experience. What we conceive and perceive will influentially infect what we do and how we do. In testing and measurement theory, according to our responses to different items – which can either be from academic achievement tests or psychological and mental tests, researchers are able to analyze how human mind works and the characteristics of a person.

The characteristics of a person are measures of hypothetical constructs. The MIRT models relate the probability of a response to person characteristics rather than to the response itself. Since the models relate probabilities of getting test items right to the characteristics of persons, they are named as "item response theory" (IRT) models. These MIRT models are different from other IRT models in that they assume there are multiple hypothetical constructs influence the performances on test items instead of only one hypothetical construct (Reckase, 2009, p. 59). The basic form of MIRT models is

$$P(U = u|\theta) = f(u, \boldsymbol{\theta}, \boldsymbol{\gamma}) , \quad (7)$$

where U is the score on the test item for a person, u represents the value of possible scores assigned to that person given the test items, $\boldsymbol{\theta}$ is a vector of parameters describing the location of the person in the multidimensional space, and $\boldsymbol{\gamma}$ is the vector of parameters describing the characteristics of the test item.

In this research, the models for the analysis are the multidimensional 2PL compensatory model and multidimensional generalized partial credit model. The multidimensional 2PL compensatory model was used to calibrate dichotomous items. The multidimensional generalized partial credit model was applied to estimate the parameters of polytomous items.

2.4.1 Multidimensional 2PL Compensatory Model

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i\right)=\frac{e^{\boldsymbol{a_i\theta_j'}+d_i}}{1+e^{\boldsymbol{a_i\theta_j'}+d_i}} \tag{8}$$

P is the probability of correct response. U is the response string with 0 indicating incorrect and 1 for correct responses. $\boldsymbol{\theta}$ is a vector of people's abilities indicating the number of dimensions in the coordinate space, since there are multiple hypothetical constructs. $\boldsymbol{a}$ is a vector of item discrimination parameters and d is an intercept term, a scalar. i is a subscript of item and j is a subscript of people.

2.4.2 Multidimensional Generalized partial Credit Model

$$\mathrm{P}\left(u_{ij}=\mathrm{k}\middle|\theta_j\right)=\frac{e^{k\boldsymbol{a_i\theta_j'}-\sum_{u=0}^{\boldsymbol{k}}\boldsymbol{\beta_{iu}}}}{\sum_{v=0}^{K_i}e^{v\boldsymbol{a_i\theta_j'}-\sum_{u=0}^{\boldsymbol{v}}\boldsymbol{\beta_{iu}}}} \tag{9}$$

K_i defines the maximum score for Item i. The lowest score is 0. There are K_i+1 score categories. The score to a person on the item is represented by $k = 0, 1, 2, \ldots, K_i$. β_{iu} is the

threshold parameter for score category u and β_{i0} is defined to be 0. All the other symbols were defined previously.

2.4.3 Statistical Description of Item and Test Function in MIRT

In UIRT, the item difficulty parameter indicates the value on the θ scale that corresponds to the point of the steepest slope for the item characteristic curve (ICC). The discrimination parameter is related to the slope of the ICC at the steepest point. The interpretation of difficulty and discrimination parameters in UIRT can be generalized to the MIRT.

In MIRT, we use item response surface to represent the relationship between multidimensional abilities and the probability of getting the item right because the location of a person is no longer on a unidimensional scale – not on a line but in a multidimensional space. The slope of a surface changes with the direction of the movement along that surface. Therefore, the point where the steepest slope is in a surface is determined by the direction of the movement.

Reckase (2009) points out that at each point in the θ – space, there is a direction indicating the maximum slope at that particular θ location. Suppose the entire multidimensional space is considered and the slopes in all directions at each point are evaluated. Then, there exists a maximum slope overall for a particular item. But, this is only true for a compensatory model. It is not true for all possible surfaces. Thus, the value of that maximum slope can be taken as a best measure of the capabilities of a test item for distinguishing between $\boldsymbol{\theta}$-points along the direction of the greatest slope.

The difficulty parameter "b" in UIRT shows the distance between the origin "0" of the unidimensional scale to the point on the θ – scale corresponding to the location of the point of steepest slope of ICC. The sign of the b – parameter indicates the direction from the 0-point to

the θ point. The negative sign means the distance is to the left of the 0 – point while the positive sign indicates the right side. The larger the b – parameter is, the harder the item is.

A similar conceptualization for the difficulty parameter can be developed for MIRT models. In MIRT, the multidimensional difficulty is the distance from the origin of the θ-space (usually the **0**-vector) to the $\boldsymbol{\theta}$-point that is below the point of the steepest slope for the multidimensional surface. The associated sign indicates the relative position of the $\boldsymbol{\theta}$-point to the origin of the $\boldsymbol{\theta}$-space (Reckase, 2009, p. 114).

Our goal is to use these statistics to determine the point where the steepest slope is for the surface in the $\boldsymbol{\theta}$-space and the distance from the origin **0**-vector to that point. The idea of introducing the distance from the origin to the steepest slope point is helpful to reparametrize the measure of item difficulty in MIRT. Reckase (2009) proposed a different representation of each point in the $\boldsymbol{\theta}$-space. Instead of using a vector of $\boldsymbol{\theta}$-coordinates, each point can be represented by a vector of angles from each coordinate axis and a distance from the origin.

In an m-dimensional space, $m-1$ angles can be computed from the θ-coordinates using trigonometric relationships. The last mth angle can automatically be calculated since the sum of squared cosines must equal to 1. The relationship between the distance from the origin, the directions from the axes, and the coordinates of the $\boldsymbol{\theta}$-point on each of the axes can be represented by the following formula:

$$\theta_v = \zeta \cos \alpha_v \,, \tag{10}$$

where θ_v is the value of the coordinate of the $\boldsymbol{\theta}$-point on dimension v, ζ is the distance from the origin to the point, and α_v indicates the direction from the axes, which is the angle between the vth axis and the line from the origin to the point.

If we substitute the exponent part in compensatory multidimensional 2PL model with the above direction and coordinate formula, then the multidimensional 2PL model can be given by

$$P(U_{ij} = 1|\zeta_j, \boldsymbol{a_i}, \boldsymbol{\alpha_j}, d_i) = \frac{e^{(\zeta_j \sum_{l=1}^{m} a_{il} \cos \alpha_{jl}) + d_i}}{1 + e^{(\zeta_j \sum_{l=1}^{m} a_{il} \cos \alpha_{jl}) + d_i}}, \quad (11)$$

where ζ_j is a scalar parameter for Person j indicating the distance from the origin to the location of the person, α_j is the vector of angles between the coordinate axes and the line extending from the $\boldsymbol{\theta}$-point to the origin in the multidimensional space. The $\boldsymbol{\theta}$-point represents the location of Person j in the $\boldsymbol{\theta}$-space. Note, the vector α_j has m elements, but only m-1 of them need to be estimated to make the sum of squared cosines equal to 1.

To find the steepest slope of the surface along the direction specified by the angle-vector α_j in the $\boldsymbol{\theta}$-space, we first need to take the partial derivative of the equation 11 with respect to ζ_j to obtain the slope expression. The partial derivative is given in

$$\frac{\partial P(U_{ij}=1|\zeta_j, \boldsymbol{a_i}, \boldsymbol{\alpha_j}, d_i)}{\partial \zeta_j^2} = P_{ij} Q_{ij} \sum_{l=1}^{m} a_{il} \cos \alpha_{jl} \quad (12)$$

From equation 12 we can find that the slope of the item response surface in the direction specified by α_j is dependent on the probability of getting the item correct, the a-parameter, and the angles with axes indicated by $\boldsymbol{\alpha}$-vector. If the angle with an axis is equal to 0°, then its $\cos \alpha_{jl} = 1$ while all other cosines are 0. Thus, the slope associated with that particular coordinate axis l simplifies to $P_{ij} Q_{ij} a_{il}$. This is the slope for unidimensional 2PL model, since this item only measures the dimension which is exactly the same as coordinate axis l when the angle between the axis l and the origin to that point is equal to 0.

To calculate the steepest slope in the direction specified by $\boldsymbol{\alpha}$-vector, we need to take the second derivative of the item response function defined in Equation 11 with respect to ζ_j and then solve for the value of ζ_j when setting its second derivative to 0. The second derivative is given by

$$\frac{\partial P(U_{ij}=1|\zeta_j,\boldsymbol{a_i},\boldsymbol{\alpha_j},d_i)}{\partial \zeta_j^2} = \left(\sum_{l=1}^{m} a_{il}\cos\alpha_{jl}\right)^2 P_{ij}\left(1-3P_{ij}+2P_{ij}^2\right), \quad (13)$$

Setting the equation equal to 0 results in three solutions. Among the three, only one will have finite value of ζ_j. That is when probability *P* is equal to 0.5 and the exponent part in item response function $\left(\zeta_j \sum_{l=1}^{m} a_{il}\cos\alpha_{jl}\right) + d_i$ is equal to 0. Thus, ζ_j is

$$\zeta_j = \frac{-d_i}{\sum_{l=1}^{m} a_{il}\cos\alpha_{jl}} \quad (14)$$

It indicates the distance from the origin to the θ-location in the θ-space where item response surface has the maximum slope along the direction specified by $\boldsymbol{\alpha}$-vector (Reckase, 2009, p. 116). From equation 14 we could also derive the formula showing the location of the point of maximum slope along a particular axis l is $\frac{-d}{a_{il}}$ because all of the cosines will be 0 except for the axis l.

The value of the slope at the steepest point in the item response surface along the direction specified by $\boldsymbol{\alpha}$-vector is

$$\frac{1}{4}\sum_{l=1}^{m} a_{il}\cos\alpha_{jl} \quad , \quad (15)$$

where the probability of getting the item correct is equal to 0.5 as shown in Equation 11. The direction of slope can be obtained by taking the first derivative of the slope equation. Thus, to solve for the direction of the steepest slope from the origin of the $\boldsymbol{\theta}$-space, we need to differentiate the equation 15 with respect to $\cos\alpha$ and set it as 0. The calculation must be solved

under the constraint that the sum of the squared cosines is equal to 1. The result is given in Formula 16

$$a_{il} - a_{im}\frac{\cos\alpha_{il}}{\cos\alpha_{im}} = 0, \quad \text{for } l = 1,2,\dots,m-1 \tag{16}$$

where $cos^2\alpha_{im} = 1 - \sum_{k=1}^{m-1} cos^2\alpha_{ik}$. Thus, the cosine of the corresponding angles with the steepest slope are given by

$$\cos\alpha_{il} = \frac{a_{il}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \tag{17}$$

Taking the arccosine of the $\cos\alpha_{il}$, we get the angles that indicate the direction from the origin of the $\boldsymbol{\theta}$-space to the point having the greatest slope considering all possible directions. These angles and cosines are characteristics of the item in MIRT (Reckase, 2009, p. 117). The cosines can be referred to as direction cosines.

Furthermore, the distance from the origin to the point of steepest slope in the direction specified in formula 17 can be presented by

$$B_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}}, \tag{18}$$

where B_i is the multidimensional difficulty of the test item. Also, the multidimensional discrimination parameter for Item *i* can be represented by

$$A_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2} \tag{19}$$

Therefore, the relation between the multidimensional difficulty and discrimination parameters is

$$B_i = \frac{-d_i}{A_i}. \tag{20}$$

2.4.4 Reference Composite

Reference composite is one of the ways to summarize the set of item characteristics in MIRT. The basic idea is to use a unidimensional scale to represent the test items in a multidimensional space. The orientation of this line indicates the direction of best measurement underlying the performance on the set of items. "This is the line in the multidimensional space that represents the unidimensional scale (Reckase, 2009, p. 126)." If we apply a unidimensional IRT model to analyze the item-response matrix with multidimensional structure, then the obtained estimates are the projections of the multidimensional θ-points on that unidimensional θ-scale in the multidimensional space.

Wang (1985, 1986) proved that the unidimensional θ-line represents the θ-estimates related to the characteristics of the matrix of the discrimination parameter $\boldsymbol{a}$ for the multidimensional compensatory model. Here, the orientation of the reference composite is given by the eigenvector of the $\boldsymbol{a}'\boldsymbol{a}$ matrix corresponding with the largest eigenvalue of that matrix. Note that the sum of the squared elements of the eigenvector is equal to 1. Therefore, the elements of the eigenvector can be considered as the direction cosines. These direction cosines indicates the orientation of the reference composite and the coordinate axes of the θ-space. The angle between the reference composite and the coordinate axes can be derived by taking the arccosine of the elements of the eigenvector. According to Reckase (2009), the reference composite tends to be oriented along the direction of the steepest slope from the origin of the multidimensional θ-space regarding the test characteristic surface.

2.4.5 Analyzing the Structure of Test Data

One common application of MIRT is to analyze the dimension structure of an item response data set. There is a long history of research approaches on determining the number of dimensions to capture the correlations among the data set using factor analysis. Major studies having carried forward to this day that can be performed on MIRT analysis are summarized in the following paragraphs.

Holzinger and Harman (1941, pp. 64-68) provided the expression for determining the number of variables needed to support the estimation of the factor loadings for m independent factors. The mathematical formula is

$$n \geq \frac{(2m+1)+\sqrt{8m+1}}{2}, \tag{18}$$

where m is the number of factors and n is the number of variables needed to support the estimation of the *m* factor loadings. The formula is under the assumption that there is no error in the estimation of correlations. Thurstone (1947) recommended that the number of variables needed for a plausible analysis with m factors should be "two or three times greater" than that number. He also specified a principle of factor analysis that has been influential over the years on MIRT analyses. "The scientific problem is to account for as much as possible of the relations between tests by the smallest possible number of common factors" (Thurstone, 1947, p. 82).

However, the early work of factor analysis concentrated much more on stipulating the amount of data for an analysis with a specified number of dimensions than the solution to determining the number of dimensions needed to model the data (Reckase, 2009). As computer technology improved, solving the issue of number of dimensions was achievable. Reise et al. (2000) summarized the research on determination of number of dimensions by concluding that it is better to overestimate the number of dimensions than to underestimate them. They suggested

scree plots, parallel analysis and analysis of residual correlation matrix are as good as more elaborate procedures for specifying the dimensions needed to model a matrix of test data.

Reckase (2009) clarifies there is no true number of factors, but through different statistical and psychometric methods can researchers identify a sufficient number of dimensions needed to accurately represent the major relationships in the item response data. He states too few dimensions will result in projecting the complexities of the relationship into a smaller space than is sufficient to represent the relationships. If the number of dimensions used to model an item response matrix is minimized on purpose to only diagnose the major constructs of the data matrix, then the meaning of constructs will be confusing rather than illuminating. According to Reckase (2009, p. 182), this can occur because the locations of people and items are projected from a high multidimensional space onto a lower dimensional space, which makes person locations seem close when they are not and items seem to be sensitive along the same dimensions when they are not. These situations happen because person locations or item vectors may be projected on top of each other when there are really large distances between them in a higher dimensional space.

Moreover, "the number of dimensions needed to accurately model the relationships in the item response matrix is dependent on two aspects of the data collection process – the number of dimensions on which people taking the test differ and the number of dimensions on which test items are sensitive to differences" (Reckase, 2009, p. 182). If a group of examinees are carefully selected to be capable on only one dimension, then even the items are designed to measure more than one constructs the total item response matrix can represent the differences on only one dimension. The same thing happens if the set of test items are designed to measure only one of the dimensions of variability of the examinee population, the data matrix will be analyzed to be

well fit by a model with a single ability θ no matter how many abilities the examinees may capture. Therefore, Reckase (2009) points out that "the number of dimensions needed to accurately model a matrix of item response data is the smaller of the dimensions or variation for the people and the dimensions of sensitivity for the test items (p. 182)."

A prominent implication setting the stage of this research is that the dimensions of variability not only depends on the set of items but also on the sample of examinee groups. The number of dimensions needed to model the variations among the item response data will be different given the different sets of items and different examinee groups. The multidimensional nature of test data is determined by the variations from both items and examinees (Reckase, 2009).

However, few of the current research studies on reporting subscores undertake this premise. Most research do not check the dimensionality analysis and assume that the number of subscore category classified by item writers or content developer is appropriate and correct. Such an assumption does not reflect the multidimensional nature of the whole data set but the test design only from content developers' perspective. If the number of dimensions is poorly-defined and the items clustered in one dimension actually do not measure the particular content those item are designed to measure, then the subscores reported on these clusters cannot be trustable even though the statistical methods to calculate the subscores could be reasonable.

When reporting subscores using latent variable θ, most researchers would apply MIRT models, higher-order IRT models, or even cognitive diagnostic models. All these models require the identification of multiple constructs/dimensions/abilities/clusters. If the classification of items are inappropriate, then the results will be misleading. When applying unidimensional models, we do not have such concerns since there is one and only one underlying construct/latent

variable/dimension/cluster. However, when transferring from a unidimensional scale to a multidimensional space, the valid and meaningful dimensionality analysis is required.

2.5 Procedures for Determining the Required Number of Dimensions

2.5.1 Parallel Analysis

Parallel analysis is one approach for determining the number of dimensions needed to describe the interactions between variables in a data set (Reckase, 2009). It has been described by Ledesma and Valero-Mora (2007) and others. It has a long history in the factor analysis literature when Horn (1965) first mentioned it in the last century. Parallel analysis has two procedure. The first step is to apply a dimensional analysis by performing the eigenvalue decomposition based on the inter-item correlations of the item-score matrix. After obtaining the eigenvalues and eigenvectors, the eigenvalues can be plotted against the number of dimensions extracted from the data matrix using a traditional scree plot. The eigenvalue decomposition was programmed using the default operation in Matlab (2015a).

The second step is to simulate a set of test data that has no relationship among the items but has the same proportion correct for each item as the real data and the same sample size. The simulated data set is then analyzed using the same eigenvalue decomposition procedure to obtain the eigenvectors and the associated eigenvalues. Since the simulated data has the same proportion correct for items as the real data, the difficulty factors generated from the simulated data set should be the same as from the real data. Thus, the eigenvalues from the simulated data are then plotted on the same scree plot with the eigenvalues from the real data using the same number of factors. The larger number of eigenvalues from the real data is then determined. That number is the number of dimensions suggested by the parallel analysis. Ledesma and Valero-

Mora (2007) suggest replicating the procedure of simulating data from the real data with the same proportion correct and same sample size for a number of times to estimate the sampling variations of the eigenvalues.

2.5.2 Clustering Analysis

Clustering analysis is another approach to determining the number of coordinated axes needed to model the relationships in the data set (Reckase, 2009, p. 220). The number of clusters does not necessarily represent the number of dimensions needed for the data set. Instead, the number of clusters from the clustering analysis sets the upper limit of the number of coordinate axes in the multidimensional space that supports the dimensional structure in the data set. It is feasible to use fewer dimensions than clusters to model the relationships in the data matrix. The smaller set of dimensions are sufficient to represent such relationships.

There are two steps to determine the clustering of items. The first step is to measure the similarity between items. The second one is to decide for the algorithm for forming clusters. Many options are available for both of these two decisions in clustering analysis literature. In MIRT application, two major methods can be adapted for each of the two decisions. For the measure of similarity between items, Miller and Hirsch (1992) proposed to compute the angle between each pair of item arrows. The other option is to measure the conditional variance between the items.

Kim (2001) analyzed a number of clustering methods regarding the measure of similarity between items and concluded that the Ward's method (1963) could recover the underlying structure of the multidimensional data more precisely than other alternatives. The implementation of Ward's method for clustering is further explained in the following section.

Suppose the vector α_1 contains the angles between item arrow and coordinate axes for Item 1 and the angles with the coordinate axes for the item arrow for Item 2 are in vector α_2. According to Harman (1976, p. 60), the angle between Item 1 and Item 2 can be calculated by the inverse cosine of the inner product of the cosines of the angles with each of the coordinate axes for each item:

$$\alpha_{12} = arccos(cos\boldsymbol{\alpha_1'}\boldsymbol{\alpha_2}) \tag{19}$$

Note that α_{12} is a scalar value since the two lines extended from two item arrows will intersect at the origin of the multidimensional space. Also, two intersecting lines will fall within a two-dimensional plane. Therefore, only one angle can be obtained between two lines extended from two item arrows.

In MIRT literature, the Formula 19 can be substituted by the discrimination parameter $\boldsymbol{a}$ for the items. The relationship between the angles between two item arrows and the MIRT discrimination parameters $\boldsymbol{a}$ can be represented as

$$cos\alpha_{12} = \frac{\boldsymbol{a_1'}}{\sqrt{\sum_{l=1}^{m} a_{1l}^2}} \frac{\boldsymbol{a_2}}{\sqrt{\sum_{l=1}^{m} a_{2l}^2}} \tag{20}$$

According to the formula in (11), if two item arrows are aligned with each other, then the angle between them is equal to 0. If two item arrows are orthogonal to each other, then the angle between them is 90°. Therefore, as the angle increases, the items will have their corresponding directions associated with the maximum discrimination parameters in different directions in the θ-space.

Harman (1976, p.63) proves that the cosine of the angle between two variables, in MIRT test items, is the correlation between the two items. In other words, the continuous latent variables underlying the performance on two items are correlated given the angles between the items. For example, when two items point in exactly the same direction, the angle between them

is 0°. The cosine of 0° is equal to 1, indicating that the correlation between the continuous latent variables underlying the performance on these two items is 1. On the other hand, if two item arrows point at right angle to each other, then there is a zero correlation between the underlying continuous latent variables because the cosine of 90° is 0. In MIRT, the input data for clustering analysis is the matrix of angles between all possible pairs of test items (Reckase, 2009, p. 221).

When clustering items using Ward's method in MatLab, the items that have the smallest angles in-between will be clustered together. Later, the program will select items that have the second smallest angles between the items that have already been clustered and create new clusters with those old clusters. These procedures will be repeated until all the items in the data set have been grouped within appropriate clusters. Therefore, the final cluster dendrogram shows that the items measuring the similar knowledge and skills are grouped together while the items measuring distinct combination of knowledge and skills are classified into different clusters.

2.6 Transformation of Parameter Estimates between Coordinate Systems for Different Test Forms

2.6.1 Test Forms and Test Specifications

Millman and Greene (1989) defined a test form a set of test questions that is designed based on content and statistical test specifications (as cited in Kolen and Brennan, 2004, p. 2). Test specifications are the blueprints that test developers will use to "ensure that the test forms are as similar as possible to one another in content and statistical characteristics" (Kolen and Brennan, 2004, p.2). Multiple test forms are used to maintain test security. Most test forms are constructed using the same test specifications and are considered as equivalent test forms.

From each test form, an item-response matrix data set can be collected. If we run the clustering analysis for a particular test form, then there will be a specific multidimensional space – coordinate system defined for that test form. For multiple test forms, there can be different corresponding coordinate systems defined. After identifying different clusters of items in a coordinate system for each individual form, we want to compare the subscores obtained from different clusters across different test forms.

In order to compare the subscores from different clusters in different test forms, we need to transform the parameter estimates between different coordinate systems from different multidimensional spaces. These parameters are item parameters and person parameters from different multidimensional item response models using both dichotomously-and-polytomously scored items. In the IRT literature, the invariance and indeterminacy are often discussed.

For MIRT model, there are three types of indeterminacy – placement of the origin, selection of units of measurement along axes, and orientation of the axes. In all cases, the locations of the persons and the characteristics of the items are invariant. In other words, "proficiencies and other characteristics of the persons will remain constant and the sensitivity of items to differences in persons' characteristics will be the same (Reckase, 2009, p. 235)." This results in the invariance of item-person interaction which generates the same probability of response for the person to the test item.

Therefore, the invariance of the MIRT model indicates that the probability of the selected response do not change with the change in the coordinate system. The indeterminacy of MIRT model means that the results obtained from different coordinate systems should be equally good. Thus, the users need to make the decision of what origin, units, and orientation of axes are most convenient for a particular test design and measurement application (Reckase, 2009).

The invariance and indeterminacy of both UIRT and MIRT literature is commonly defined and applied in item-person calibration software. For MIRT, there is no standard coordinate system. Estimation programs will usually set their own system of convenience. One typical method is to set the origin of the solution space to have a mean $\boldsymbol{\theta}$-vector as $\mathbf{0}$-vector. Another common choice in MIRT estimation program is to use the standard deviation of a coordinate axis as the unit of measurement for that axis. Different estimation programs can set units of measurement in different ways.

A third adjustment that is commonly used in MIRT estimation program is to fix the correlation between coordinates axes to zero. Such constraint simplifies the statistical procedure for parameter estimation because it forces to place the axes of the coordinate system in a particular orientation. The zero-correlation constraint on coordinate axes also rotates the original coordinates to yield better interpretation of the constructs being assessed associated with each particular coordinate axis.

Before estimating the person and item characteristics using MIRT models, we needed to select software that could satisfy the requirements of invariance and indeterminacy of MIRT models. Chapter 3 – section 3.2.1 discusses the details of the selection of the software – FlexMIRT and Mplus for parameter estimation in this study.

2.6.2 Recovery of the Rotation and Scaling Matrices

The space we are interested in MIRT is the space defined by the location of persons. These locations are represented using $\boldsymbol{\theta}$-vectors. The origin of the $\boldsymbol{\theta}$-space, the rotation of coordinates, and the units of measurement are all arbitrarily determined by people who apply the MIRT analysis or people who write the estimation programs to estimate the model parameters.

The invariance property of the MIRT models should hold when applying the indeterminacy property. In order to fulfill these properties, the expressions for the exponents of the MIRT model and items should yield identical values. That is,

$$\boldsymbol{\upsilon\nu'}+\boldsymbol{\zeta}\mathbf{1} = \boldsymbol{a\theta'}+\mathbf{d1}, \tag{21}$$

where $\boldsymbol{\upsilon}$ is the matrix of discrimination parameters for the transformed space, $\boldsymbol{\nu}$ is the matrix or person parameters after transformation, and $\boldsymbol{\zeta}$ is the intercept parameter after transformation.

If the transformation does not involve rotation of coordinate axes or change of the unit of measurement, then the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ - the person location from the first coordinate system and the second coordinate system is given by

$$\boldsymbol{\nu}_j = \boldsymbol{\theta}_j - \boldsymbol{\delta}, \tag{22}$$

where $\boldsymbol{\nu}_j$ is the vector of coordinates for Person *j* in the new coordinate system, and $\boldsymbol{\theta}_j$ is the vector of coordinates for Person *j* in the old coordinate system, and $\boldsymbol{\delta}$ is the vector representing the location of the new origin using the old coordinate system. Therefore, the matrix equation for converting the coordinates of the person locations from the old coordinate system to the new coordinate system is

$$\boldsymbol{\nu} = \boldsymbol{\theta} - \mathbf{1}\boldsymbol{\delta}, \tag{23}$$

If the transformation includes the change of origin and unit of measurement, and the rotation of coordinate axes, then the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ is given by

$$\boldsymbol{\nu} = \boldsymbol{\theta RotC} - \mathbf{1}\boldsymbol{\delta'RotC}, \tag{24}$$

where *Rot* is the rotation matrix and *C* is the scaling matrix.

Therefore, given the above formulas of the relationship of person parameter transformation between different coordinate systems, the relationship between item parameters

before transformation to those after transformation can be determined by replacing the term of $\boldsymbol{v}$ in equation x with equation xx. Thus, the left side of equation x can be transformed into

$$\boldsymbol{v}(\boldsymbol{\theta RotC} - \mathbf{1}\boldsymbol{\delta}'\boldsymbol{RotC})'+\boldsymbol{\zeta}\mathbf{1} = \boldsymbol{v}(\boldsymbol{\theta RotC})' - (\mathbf{1}\boldsymbol{\delta}'\boldsymbol{RotC})' + \boldsymbol{\zeta}\mathbf{1}$$

$$= \boldsymbol{vC}'\boldsymbol{Rot}'\boldsymbol{\theta}' - (\mathbf{1}\boldsymbol{\delta}'\boldsymbol{RotC})' + \boldsymbol{\zeta}\mathbf{1}. \quad (25)$$

Comparing the second line in Equation 25 with the right side of Equation 24, we obtain the equivalent of $\boldsymbol{a}$ in

$$\boldsymbol{a} = \boldsymbol{v}\boldsymbol{C}'\boldsymbol{Rot}' \quad (26)$$

Since $\boldsymbol{a}$ and $\boldsymbol{v}$ can be obtained from the calibration, then the result can be solved for $\boldsymbol{C}'\boldsymbol{Rot}'$. This solution requires the nonorthogonal Procrustes procedure to determine the rotation and scaling matrices for the transformation of item parameters between two coordinate systems. In Equation 26, $\boldsymbol{a}$ is used as the target matrix before transformation while $\boldsymbol{v}$ is the new matrix after transformation. Here, the nonorthogonal Procrustes is used to determine the transformation from $\boldsymbol{v}$ to $\boldsymbol{a}$. The recovery of the rotation matrix and the scaling matrix can be represented as

$$\boldsymbol{C}'\boldsymbol{Rot}' = (\boldsymbol{v}'\boldsymbol{v})^{-1}\boldsymbol{v}'a$$

$$M=\boldsymbol{C}'\boldsymbol{Rot}' \; . \quad (27)$$

where M is the rotation and scaling matrix to transform the parameters from the new coordinate space to the old coordinate space. Once we obtain the matrix M, we can postmultiply M^{-1} by the subscores $\boldsymbol{\theta}'$ in the new coordinate system to get the recovered subscores $\underline{\boldsymbol{\theta}}$ in the old coordinate system. The formula is given as

$$\underline{\boldsymbol{\theta}}'=M^{-1}\,\boldsymbol{\theta}' \quad (28)$$

The rationale of this formula is to achieve the invariance property of MIRT model. That is the location of persons do not change and the probability of getting items correct will not change during the procedure of transformation. Thus, the exponent part - $\boldsymbol{a\theta}'$ in any MIRT

model should not be changed. When person parameter $\boldsymbol{\theta}$ is transformed using $R^{-1}\,\boldsymbol{\theta}'$, the item parameter $\boldsymbol{a}$ also needs to be transformed by the rotation and scaling matrix $\boldsymbol{M}$ to achieve the invariance property. The transformed $\boldsymbol{a}$ is given by

$$\underline{\boldsymbol{a}} = \boldsymbol{a}M \tag{29}$$

Since $\boldsymbol{MM^{-1}} = \boldsymbol{I}$, $\underline{\boldsymbol{a\theta}'} = \boldsymbol{aMM^{-1}}\,\boldsymbol{\theta}' = \boldsymbol{a\theta}'$, where $\boldsymbol{MM^{-1}} = \boldsymbol{I}$.

2.7 Linking, Scaling and Equating

Test centers report scores for different examinees taking tests on different test dates. For international tests, examinees take tests in different countries. Therefore, test scores reported for examinees taking tests in different places on different test dates should be interchangeably comparable and interpretable. Dating back to early 20th century, Kelly (1914), Starch (1913), Weiss (1914), and Pinter (1914) discussed the methods of putting scores from different tests into comparable units. It was not until decades after the invention of scaling methods that the desire to equate scores on alternative forms of the same test arose (Holland and Dorans, 2006, p196). Levine (1955) and Lord (1980) made great contributions in the application of equating test forms using classical test theory and item response theory.

2.7.1 Unidimensional Linking, Scaling and Equating

Linking is a general concept of transforming scores from one test form to another. Linking can be divided into three categories: predicting, scaling, and equating (Holland and Dorans, 2006, p. 187). Predicting is to predict an examinee's scores on one test form from other information about that examinee, such as a score on another test form, demographic information, and so forth (Holland and Dorans, 2006, p. 188). Scaling is to transform the scores from different test forms onto a common scale so that the scores can be comparable from different test forms (Holland and Dorans, 2006, p. 189). According to Kolen and Brennan (2004), "Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. […] The process of equating is used in situations where such alternate forms of a test exist and scores earned on different forms are compared to each other" (pp. 2 - 3).

2.7.2 Use Multidimensional Linking to Report Subscores across Different Test Forms

In the MIRT context, people need to know the difference of examinees' locations in the coordinate system defined by the tests at two or more times. The determination of amount of change in location requires putting item-and-person parameter estimates from two or more different test forms into the same coordinate system. The goal is to report test scores using the same reporting system (Reckase, 2009, p. 275).

2.7.3 Identifying the Common Multidimensional Space

The first step in the application of linking is to specify a base coordinate system. Our goal is to transform all the calibrated results of item and person parameters from different

coordinate systems into this base coordinate system. Thus, the subscores obtained from each individual test forms are comparable.

In order to achieve such goal, we need to keep the origin of the coordinate system, the units of measurement and the rotation of coordinate axes consistent accordingly to those in the base coordinate system. For example, one option is to use the same software to run the calibration for each test form so that the invariance and indeterminacy properties of MIRT can hold.

Therefore, the selection of the base coordinate system is very important. Instead of selecting a default output from an estimation program, we need to thoughtfully consider how to set up a base coordinate system in MIRT application. Unlike in the UIRT application, the selecting of a base coordinate system in MIRT is much more complex. Not only do we need to decide the location of the origin of the coordinate system, but also need to set the units of measurement on each coordinate axis. Moreover, the determination of the orientation of coordinate axes is also essential. That is, we should build up parallel multidimensional coordinate systems across different test forms so that we can compare the subscores from each individual test form.

When multiple test forms were built up using the unidimensional IRT model, how to report valid subscores that can be comparable across different test forms requires linking these test forms using multidimensional item response theory. The first step is to specify a common or base coordinate system. The second step is to transform the subscores from each individual form onto that base form. That is to link the test forms using multidimensional item response theory.

If we consider each test form as a data set and set up a corresponding multidimensional coordinate system for each data set, then the question here is we will get different coordinate

systems or the same coordinate systems? The coordinate system is composed of coordinate axes, the origin of the space, and units of measurement. When using the same software to estimate the item and person parameters, the origin of the space and units of measurement are default setting in the program. The only thing that is to be determined is the number of dimensions. In other words, we need to tell the estimation program how many dimensions or coordinate axes can be identified in the data set.

We can solve the number of dimensions by running the parallel analysis with 100 simulations. The number of dimensions specified in the base coordinate system is the one that needs to be defined in the calibration program. Here, additional questions should be considered: which data set can be selected to run the parallel analysis? In other words, which data set can be used to generate the base coordinate system? Is this the one that is selected among those individual test forms or is it the one that has contained enough responses to run the multidimensional analysis?

In order to answer these questions, we first need to run the dimensionality analysis to check the number of dimensions that is required to capture all the relationship within each data set. Moreover, we need to set up the data set for the base coordinate system, because all the subscores will be transformed back into this coordinate system. After obtaining the rotation and scaling matrix between each individual form and the base form, we will use these rotation matrices to link the subscores onto the same dimensions in the base space. The formula is given by

$$\underline{\boldsymbol{\theta}} = M^{-1}\boldsymbol{\theta}', \tag{30}$$

and its rationale is explained in Chapter 2 – section 2.6.2.

Once we select the data set for the base coordinate system, we will run the parallel analysis to identify the number of dimensions that is appropriate to represent the relationship among item-person response in the data set. The number of dimensions will then be used to run the calibration on the estimation of item and person parameters for each individual test form. The cluster analysis will also be applied to specify reasonable clusters for items within the data set with the number of dimensions identified from the base coordinate system. After the group of items are clustered together, we will compute the reference composite for each cluster. That is, for an individual test form, each cluster is represented by a reference composite. Finally, we will link different test forms based on the common reference composites.

2.8 Relating Subcores from Different Test Forms

Many test programs have more than one test forms to maintain test security or to control the memory of responses in pretest and posttest experimental designs. Usually, these multiple test forms are designed using the same test specification and are intended to be equivalent forms of each other. In these cases, putting the results from different test forms into the same coordinate system is called equating (Reckase, 2009).

For those test forms that are not constructed according to the same test specification, if we still need to report the results using the same coordinate system, then even though the statistical procedures are the same as are used for equating, such processes are called linking or scaling for comparability because the obtained estimates of constructs do not have the same level of accuracy and may have slightly different meaning (Reckase, 2009).

In this research, the goal is to report subscores from different test forms. Thus, the multidimensional linking and equating will be applied accordingly to achieve the research goal. But, the focus of this research is not on the application of multidimensional linking and equating.

So, the details of linking and equating will not be introduced here. There are extensive books on equating and linking of calibrations such as Dorans, Pommerich and Holland (2007), Kolen and Brennan (2004).

2.8.1 Common-Person Design

The common-person design is also named as single-group design. It means two different test forms are administered to same examinees. There are two defects regarding this design. First, the fatigue factor will influence the testing results given same group of examinees taking the same test using two different forms on different test dates. Second, the familiarity with the test information will increase the performance for the later tests. Therefore, single-group design is rarely applied in practice due to the fatigue issue and the order of effect issue.

Under multidimensional condition, the common-person design means reporting the scores from two different test forms within the common multidimensional space. Thus, the goal here is to put the ability estimates (θ-estimates) from different test forms into the same coordinate system. Such a design needs to be applied under the condition that no change has to be made to the examinees between the administrations of test forms.

2.8.2 Common-Item Design

Although the concept of common-person design is easy to understand, it is not easy to apply this design in operational linking and equating procedure. There are some concerns that for examinees who take the same tests it is impossible for them to be motivated on the same level when responding to different test forms, not to mention other carry-over effects. For example, counterbalancing the administration of different test forms may not be sufficient to balance

differences in motivation. Also, examinees may experience some fatigue when taking the same tests again and again. Therefore, other equating designs are more often implemented.

The common-item design is a frequently used method for linking and equating. The general idea is both test forms maintain a subset of common items which are identical across the forms. The size of the common items occupies 20% to 25% of the total test length. These common items are selected to span the content of the test as a whole and are designed to be as parallel as possible to the full test. That is, these common items can be considered as mini version of the full test and function exactly the same way for different test forms.

2.8.3 Randomly Equivalent-Groups Design

Another commonly used equating design is randomly equivalent-groups design. It is used when tests are assumed to be sensitive to differences in examinee performance on a single composite of skills. The examinee sample is randomly divided into two or more subsamples with the number of samples equal to the number of test forms to be equated. The assumption is that the distribution of the examinee performance on the composite skill being measured should be the same across all different test forms due to the random assignment of forms to samples. If there are any differences among these distributions of performances, then they are because of the slight differences in the characteristics of the tests. Transformation of scores from each individual test form to a common scale requires the distributions with the same features for all test forms.

The randomly equivalent-groups design will be effective once the carefully designed test forms can be assumed to result in score scales that have the same conceptual meaning. In other words, the constructs represented by the test scores are the same across all forms. Moreover, the

distributions of the performance are assumed to be the same due to the random assignment of the test forms to groups.

When applying the randomly equivalent-groups design in MIRT case, it becomes more complex than in the UIRT case because the orientation of the coordinate axes might not be the same for the calibration of the two test forms even when the examinee groups are random samples from the same population. Further, there are no common items between two test forms that can be used to determine the rotation needed to align the coordinate axes. Therefore, the application of randomly equivalent-groups design in MIRT case involves developing a criterion for rotating the coordinate axes to a common multidimensional space (Reckase, 2009).

There will be future work in MIRT for linking and equating using other methods. In this research, the solution we have is to compute the reference composite for the clusters assessed by the tests. Then, we can use those common reference composites to link test forms. That is, once the constructs can be identified for multiple test forms, then the coordinate axes can be transformed to a common orientation by solving the rotation and scaling matrix to align the reference composites from different test forms (Reckase, 2009).

One assumption needs to be satisfied when applying the above method. That is, test forms are constructed to be parallel or they measure common constructs. Such assumption not only applies for the unidimensional but also works for the multidimensional case. Therefore, before relating results or subscores from different test forms into one common orientation multidimensional coordinate system, we need to assure the multidimensional structure from different individual test forms are parallel, or at least these forms are measuring the common constructs.

The application of multidimensional linking and equating requires the common dimensionality structure and common constructs across different forms. This is the same as what is done in unidimensional linking and equating. However, most test forms are designed using unidimensional model and are linked and equated based on unidimensional structure. When reporting subscores for these test forms, we need to set up a reasonable transformation between unidimensional structure and multidimensional structure.

In order to achieve such goal, we should apply the parallel analysis, exploratory factor analysis, confirmatory factor analysis, clustering analysis, and reference composite computation to link and equate different test forms and to report and compare subscores across different test forms.

In general, the procedures for linking and equating test forms using multidimensional IRT are as follows:

Step 1. Set up the base coordinate system so that any results from individual test forms can be transformed back into this base form

Step 2. Set up the coordinate system for each individual test form so that subscores from different clusters or constructs can be computed

Step 3. Run the parallel analysis, exploratory and confirmatory analysis, and clustering analysis for the base form to identify the target matrix for the nonorthogonal rotation

Step 4. Run the parallel analysis, exploratory and confirmatory analysis, and clustering analysis for each individual form

Step 5. Calibrate the data for each individual form and determine the directions of the reference composites in individual coordinate space. The direction cosines of the reference composite are used as the $\boldsymbol{a}$- parameters for the reference composites. These $\boldsymbol{a}$- parameters will

be collected into a matrix and used to compute the nonorthogonal Procrustes rotation matrix between each individual test form and the base form

Step 6. Determine the nonorthogonal Procrustes rotation that will convert the directions of the reference composites in individual test form into the directions of the reference composite in the base form

Step 7. The rotation matrices will be used to covert the locations of examinees from each individual test form to the base form. These transformed locations of examinees for each different clusters are the subscores that can be compared across different test forms.

CHAPTER 3
METHODOLOGY AND RESULTS

3.1 Data Description and Data Analysis Procedure

The purpose of the test used in this research is to measure non-native-English speakers' ability to function in an academic environment where instruction is provided in English. Specifically, this test assesses proficiency in listening, reading, speaking, and writing English. It uses 20 different item types that assess different communicative skills, enabling skills and other traits. Detailed descriptions of the item types are provided in Pearson Longman (2010).

This study used data from 36,938 examinees, 954 items, and 164 test forms from over 165 countries. Those with the largest number of examinees included China, India, the United States, Japan, South Korea, Australia, the United Kingdom, Hong Kong, Taiwan and Canada. Unfortunately, even though this is a large data set, the number of examinees responding to each test form was lower than desired for stable estimation of the parameters of a MIRT model. Therefore, individual form data were used to check the generalizability of results obtained from a large set of common items across forms. The large set of common items was used to identify an overall dimensional structure that was checked against the dimensional structure of individual forms.

In order to have sufficient data for stable estimation of MIRT model parameters, the most frequently used 100 items over all test forms were selected for analysis. One problem with this approach was that the most frequently used 100 items did not have the same distribution over item types as a full test form. The use of the most frequently used 100 items had both advantages and disadvantages. The advantage was getting very stable estimates of model parameters and good evidence of the dimensional structure of the item types that were present. Often there were numerous items of a particular type in this data set. The disadvantage was that

the results from the analysis might not represent results to be expected from operational test forms. For that reason, the results obtained for the most frequently used 100 items were checked with analyses of the four most frequently used test forms.

Of the 164 test forms, four were found to have sufficient data for the multidimensional analyses. The minimum sample size for the forms was 432. Thus, the analysis data consisted of five data sets. The first data set is the 100 items with highest frequencies of use. This was used to obtain results that could generalize across all test forms. The second to fifth data sets are from the four test forms with highest frequencies of administration. These were used to confirm the results from the 100 most frequently used items and to check the consistency of findings across forms.

Table 1 shows the distributions of item types for a test form. Table 2 provides the number of examinees and the number of items for the five data sets. Table 3 shows the number of common items between pairs of the five analysis data files. These common items are from different content categories labelled by the content experts during the item writing and test development. Therefore, these common items are distributed over different content categories across different test forms.

TABLE 1:

Item Types and Content Distribution for One Form of the Test

Part and Section	Item Types	Number of Items
Part 1: Speaking	Read aloud	6
	Repeat sentence	10
	Describe image	6
	Re-tell lecture	3
	Answer short question	10
Part 2: Writing	Summarize written text	2
	Write essay	2
Part 3: Reading	Multiple-choice, choose single answer	2
	Multiple-choice, choose multiple	2
	Re-order paragraphs	2
	Reading: Fill in the blanks	4
	Reading and writing: Fill in the	5
Part 4: Listening	Summarize spoken text	2
	Multiple-choice, choose multiple	2
	Fill in the blanks	2
	Highlight correct summary	2
	Multiple-choice, choose single answer	2
	Select missing word	2
	Highlight incorrect words	2
	Write from dictation	3

TABLE 2:

Number of Examinees and Number of Items for the Five Analysis Data Sets

Data Sets	Number of Examinees	Number of Items
Dataset 1 100 Items with Highest Frequencies	36938	100
Dataset 2 Form 1	448	65
Dataset 3 Form 2	438	53
Dataset 4 Form 3	437	69
Dataset 5 Form 4	432	66

TABLE 3:

Common Items between Pairs of the 100 Items and Four Test Forms

	100 Items	F1	F2	F3
F1	16			
F2	14	3		
F3	25	1	6	
F4	21	21	0	2

3.2 Dimensionality Analysis

Parallel analysis, exploratory and confirmatory factor analysis, cluster analysis and reference composite analysis were used to investigate the structure of the dimensions among the five data sets - most frequently used items and most frequently used forms. The procedures are described in detail in the following sections.

3.2.1 Software Check – FlexMIRT VS. Mplus

Due to the complexity in the data sets – a large sample size, a large group of mixed item types composed of both dichotomous and polytomous items, and missing values, the estimation capability of software is essential to this research. The calibration results using MIRT models have to be accurate and reliable. Otherwise, the results will not be valid and meaningful. The major software selected for this research is FlexMIRT (2013, Li et. al). The reason is that it can calibrate item parameters and multidimensional person-abilities for polytomous MIRT models (Houts and Cai, 2013). Mplus was also used to run the same analyses for quality assurance purpose. Interested readers can refer to the paper (Reckase and Xu, 2015) for the research design and its corresponding results obtained from Mplus.

Before running the analyses on real data sets, we first checked the quality and efficiency of FlexMIRT. In order to do that, we simulated the data sets with exactly the same information and the format as the real data sets. That is, the simulated data sets contained both dichotomous and polytomous items with large amount of missing data, and same number of examinees. Then, we ran the FlexMIRT under different conditions 1) dichotomous items 2) polytomous items 3) different number of examinees. Later, we compared the results obtained from FlexMIRT with true values generated in simulation and those from Mplus as well using the same simulated data sets to double check whether FlexMIRT could be used to run the calibration on the real complex data sets appropriately. Finally, after confirming the quality and validity of the software, we applied FlexMIRT to run calibrations on all five real data sets using the multidimensional generalized partial credit model.

3.2.2 Simulated Data Sets

We first simulated eight clusters within a six-dimension multidimensional space. The six dimensions are the results obtained from Mplus. The parallel analysis based on the F100 data set showed that both seven-dimension and eight-dimension solutions were efficient enough to capture the correlations among the data set. Then, we ran cluster analysis using seven and eight dimensions respectively. Both results returned six clusters for F100. When determining the number of dimensions in a data set, the conservative number of dimensions is preferred. Therefore, we selected seven-dimension structure as the base space.

Eight clusters were obtained from parallel analysis using Matlab given the seven-dimension structure. Before the analysis of mixed structure of dichotomous and polytomous items, we first simulated a data set with 10,000 examinees and 100 items. The examinees were randomly selected from multidimensional standard normal distribution with seven dimensions. All item responses were dichotomously generated using MIRT 2PL model.

Next, we ran the parallel analysis using MatLab to check the number of dimensions for the simulated data set. The results from parallel analysis showed that the simulated data set had six dimensions. It had less dimensions because the simulated data set had less examinees and only dichotomous items. Then, we used FlexMIRT and Mplus to run exploratory factor analysis separately by defining six dimensions. After obtaining discrimination a-parameters associated with each dimension for all 100 items, we used MatLab to run cluster analysis on the two a-parameter matrices. One was from FlexMIRT analysis and the other was from Mplus analysis. We did this to compare the efficiency and workability between the two software. The cluster analysis results showed that both FlexMIRT and Mplus returned eight clusters. The results from

the simulated data set indicated that FlexMIRT worked well on dichotomous items using the MIRT 2PL model.

Later, we expanded the number of examinees from 10,000 to 20,000 but still simulated the same number of 100 items and eight clusters within six dimensions. For real data set, there are 36,938 examinees. So, we needed to check whether both software worked well for large sample of examinees. We did the analysis following the same procedure – parallel analysis, exploratory factor analysis, and cluster analysis. The results indicated that both FlexMIRT and Mplus worked well for large sample size.

Finally, we simulated the data sets by adding the missing values that have exactly the same pattern as those in the real data set for the first 20,000 examinees and 100 items in F100. We ran the parallel analysis, EFA, and clustering analysis using both FlexMIRT and Mplus. The results showed that FlexMIRT and Mplus could handle missing values very well.

Moreover, FlexMIRT provided more accurate results since it used the MIRT 2PL model to calibrate the item parameters directly while Mplus did not. The a-parameters obtained from FlexMIRT are linear transformation of factor loadings obtained from Mplus. Therefore, the factor loadings could also be used to check for the dimensionality and clustering of items in data sets.

3.3 Real Data Sets

3.3.1 Analysis of Most Frequently Used Items – Form F100

We first ran the parallel analysis to determine the number of dimensions needed to capture the correlations (variations) among the data set. The number of dimensions identified can be used to set up the coordinate axes for the common space since they are stable and

consistent. Next, we re-scored the items in the F100 for the first 10,000 examinees. We examined the score distribution for each item to decide which score categories can be combined together. Such procedure was named as "repolytomizing" in the following sections. The reason to do so was to have enough examinees for each score category of polytomous items so that the item parameters could be calibrated using the multidimensional Graded Response Model appropriately.

After repolytomizing these items, we ran EFA using FlexMIRT with seven dimensions. The results of the calibration were plausible. So, we continued to repolytomize F100 but using all 36, 938 examinees. We ran EFA using FlexMIRT in a seven-dimensional space. We also ran clustering analysis, but the results were not as good as expected. The classification of clusters was not distinguishable. Therefore, we redesigned the factor analysis to combine both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) together – CEFA.

The following section describes each step in detail.

In this study, the first step consisted of computing the eigenvalues from the inter-item correlations among the 100 most frequently used items. This was done pair-wise to account for missing data across forms resulting in a different sample size for the correlations of each pair of items. This can influence the results of the dimensional analysis.

In order to solve the missing values due to the random assignment of test forms, after computing the eigenvalues from the inter-item correlations among the 100 items, we generated a data set that has the same proportion item correct with the missing values replacing exact the same places as in the real data for the 100 items. Multinomial model was applied when generating the proportion correct in Matlab. Then, the eigenvalues from the generated data set were extracted from the inter-item correlations with 100 replications. In order to deal with the

missing values, we used the multiple linear correlation method to impute the missing values in the covariance matrix from the data set by estimating the missing values with the first five columns of the covariance matrix. In other words, we replaced the missing-value correlation labeled as NaN (stands for Not a Number in Matlab) with a real number that was predicted from the previous five columns of the covariance matrix without missing values so that all the missing values in the whole covariance matrix were approximated.

Therefore, for comparison purpose, random data sets were generated with the same proportion of item scores for each item, individual item scores were removed to exactly match with the pattern of missing values in the real data sets. Then, the eigenvalues from the generated data set were extracted from the inter-item correlations. This process was replicated 100 times to yield distributions of the eigenvalues from the randomly generated data sets.

The parallel analysis was applied based on two different statistical assumptions. The null hypothesis is there are no eigenvalues from the observed data set larger than those from the randomly simulated data set. The alternative hypothesis is there did exist number of eigenvalues from the observed data set larger than those from the randomly simulated data set. These two statistical hypotheses were established based on the fact that we developed an empirical sampling distribution over the 100 randomly generated data sets to investigate the point when the real eigenvalues are statistically significantly higher than those from the simulated distribution.

Because of the pattern of missing data, there were cases where a correlation could not be computed between a pair of items. Without the full correlation matrix, the eigenvalues could not be computed so the missing values were imputed by predicting the missing values using the data from all of the other columns of the correlation matrix using multiple linear regression. Because the same procedure was used for the real data and the randomly generated data, any artifacts

caused by the imputation and the pattern of missingness would be present in both types of data sets.

Figure 1(a) shows a plot of the magnitude of the eigenvalues from the real data and those from the 100 replications of the randomly generated data. Because the eigenvalues from the generated data showed little variation, the results where the curves cross is magnified and presented in Figure 1(b) so the number of eigenvalues from the real data that are greater than those from the random data can be identified.

Figures 1(a) and 1(b) indicated that the first eight eigenvalues for the real data were larger than the first eight eigenvalues for the random data, although the difference between the seventh and eighth real eigenvalues was very small. According to the rule suggested by Ledesma and Valero-Mora (2007), the number of dimensions needed to model the data is the number of eigenvalues that are greater than those from the random data. In this case, both seven and eight dimensions were investigated further and there was little difference in the results so the more parsimonious seven dimensions were selected for further analysis.

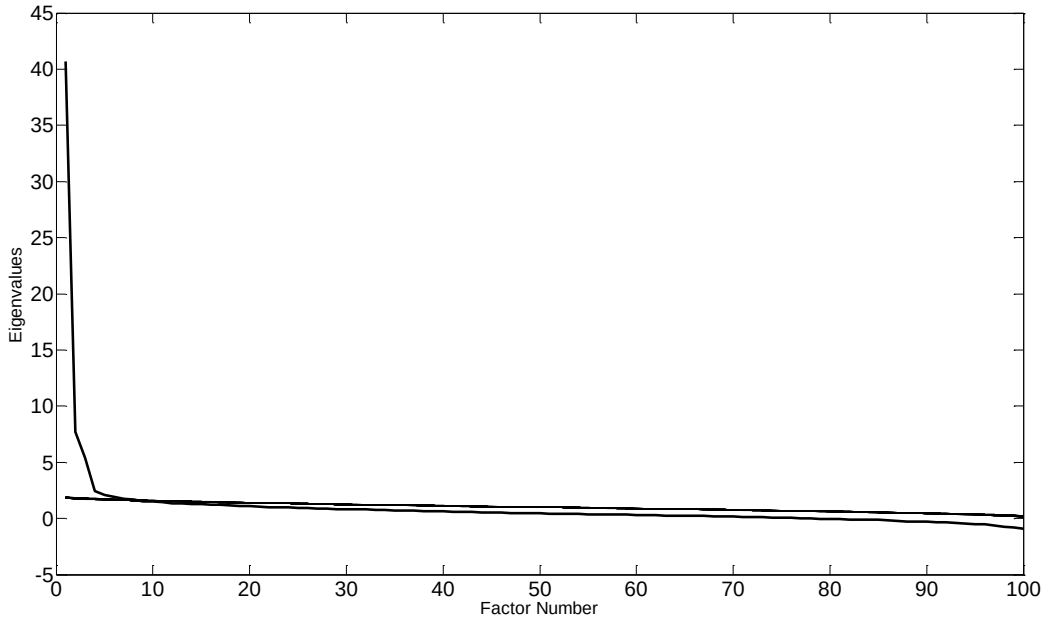


Figure 1. (a): Plot of the Eigenvalues for the Real Data and 100 Replications of Random Data

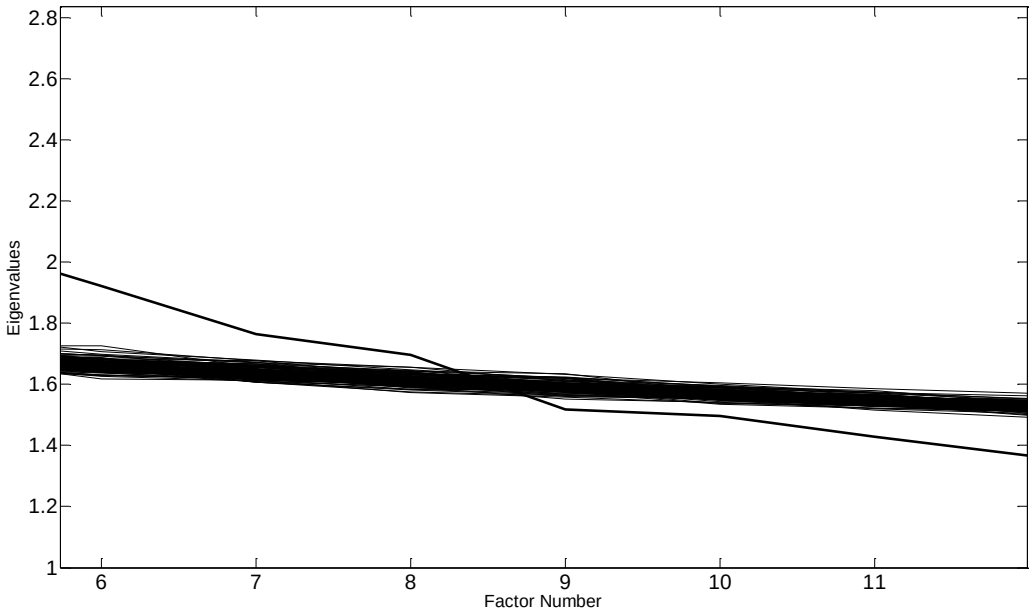


Figure 1. (b): Magnified Plot of the Number of Eigenvalues Larger than Random Data

3.3.2 Dichotomized Data Sets

The results from the simulation study showed that FlexMIRT gave good results for MIRT calibration. The next step was to use the real data sets for analysis. The F100 has 36,938 examinees and most items are polytomous items with number of score categories up to 12. So, the major differences between the simulated data and the real data are that real data have a larger sample size with more missing values and most items are polytomously scored. Our goal was to investigate whether FlexMIRT could recover a-parameters of dichotomous items in real data using MIRT 2PL models. The first 10,000 examinees and first 100 items from F100 with highest frequencies across all examinees and all test forms were selected, for they contained least missing values and were stable for the MIRT analysis.

First, we dichotomized polytomous items by computing the median score for all 100 items across 10,000 examinees. Next, we compared each examinee's score response with the corresponding median score for that particular item. If the score for that item was larger than the median score, then 1 was assigned; otherwise, 0 was assigned. We dichotomized all the responses to get a balanced proportion of score categories so that those polytomous score categories with small frequencies would not affect the item parameter calibration.

After dichotomizing all 100 items, we ran an exploratory factor analysis (EFA) on the dichotomized 10,000-by-100 matrix with missing values represented by -1 in a seven dimensional space. The FlexMIRT calibrated the a-parameters using MIRT 2PL model. Cluster analysis was later applied using the calibrated a-parameter matrix. There were still some convergence problems, such as a-values and d-values for some items were very large or even negative and the standard errors could not be computed. The cluster analysis showed that items

were clustered together with same item types. Therefore, FlexMIRT was supported for real data analysis with dichotomous items.

3.3.3 Polytomous Data Sets

FlexMIRT worked well for calibration of the dichotomized items with frequent missing values using the multidimensional 2PL model. The next step was to further check whether it could calibrate the polytomous items using multidimensional generalized partial credit model. We first collapsed the original polytomous score categories into smaller numbers of categories by combining score categories together where there were small counts of responses. After reducing the category number for all the items for the first 10,000 examinees in F100, we ran EFA in a seven-dimensional space using the multidimensional generalized partial credit model. The a-parameters associated with each dimension for individual items from the first 10,000 examinees among 36,938 were reasonably good even though there were few items that were not calibrated appropriately.

3.3.4 Exploratory Factor Analysis

Finally, we repolytomized all 100 items in F100 using all 36,938 examinees. In order to check the viability of using seven dimensions to compute the item parameters especially the discrimination parameter matrix – *a* matrix associated with each dimension, exploratory factor analyses were run on the 100 item dataset using FlexMIRT (2013, Li et. al). We ran EFA in a seven-dimensional space using both multidimensional 2PL model and multidimensional generalized partial credit model.

The results showed that the seven-dimensional solution gave the best combination of distinctly defined dimensions. Most items had reasonable a-parameters while only few did not. The inappropriately calibrated a-parameters might have been due to the inappropriate repolytomizing of polytomous items. Therefore, we scrutinized these items and repolytomized them again details described in Section 3.3.1. Then, we re-ran the EFA using the revised data set. We kept repeating these procedures until all a-parameters from FlexMIRT calibration were reasonable.

The results from EFA using FlexMIRT for the first 100 item data set supported the seven-dimension structure. However, these results are based on the linear regression imputation of missing data and the repolytomizing of multi-score-category items for the first 100 items. To further check the meaningfulness of the seven dimensional structure, cluster analysis procedures were used to determine if sets of items had theoretically supportable connections to the content structure of the tests.

3.3.5 Hierarchical Cluster Analysis

Within the context of multidimensional item response theory, hierarchical cluster analysis (HCA) is an approach for identifying sets of items that are best at measuring the same combination of skills and knowledge. There are two steps in cluster analysis procedure. The first step is to select a method to measure the similarity between items. The second step is to sort the items that share similarities into clusters (Reckase, 2009).

When using FlexMIRT to analyze the data sets, the $\boldsymbol{a}$-parameters from the seven-dimension solution were used as the item discrimination parameters for the multidimensional item response theory model. The average distance algorithm in the hierarchical clustering

routine within MatLab was used for the clustering. The cluster results indicate that six distinct clusters were identified through the analysis of the 100 item data set. Moreover, among these six clusters, five distinct clusters consist of unique collections of item types and one cluster is composed of a mix of three different item types. These six major clusters were labeled according to the conceptual representation of factors in the language ability domain defined by Carroll (1993, p.147). They are: (1) Cloze, (2) Listening, Oral Production (3) Listening, Writing (4) Oral Production, (5) Phonetic Coding, Spelling, and 6) Pronunciation, Word Recognition.

3.3.6 Reference Composites

The reference composite for a set of test items is a mathematical derivation of the line in the multidimensional space that represents the unidimensional scale defined by a set of items. This scale is the one that would be obtained if the items were analyzed using a unidimensional item response theory model. Wang (1985, 1986) showed that the eigenvector that corresponds to the largest eigenvalue of the $\boldsymbol{a}'\boldsymbol{a}$-matrix gives the orientation of the reference composite line in the multidimensional space. The $\boldsymbol{a}$-matrix in this case is the matrix of item discrimination parameters for the multidimensional item response theory model from the seven-dimensional solution. Because the sum of the squared elements of the eigenvector is equal to 1, the elements of that eigenvector can be considered as the direction cosines for the line representing the scale.

The reference composites were computed for each of the clusters of items identified by the cluster analysis procedure. The reference composites represent the distinct subscores that can be supported by the set of items. One way of computing the subscores is to project the estimates of locations of the examinees in the seven-dimensional space onto these reference composite lines. See Reckase (2009, p. 301) for the details of the projection method.

In order to compute the reference composite for all the items within each cluster, the ***a'a***-matrix was obtained and decomposed into eigenvalues and eigenvectors. Table 4 gives the angles in degrees between each reference composite line and the coordinate axes in seven-dimensional space for each cluster of the 100 item set.

TABLE 4:

Angles between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Six Clusters in Form F100

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7
Cloze	45.69	89.52	79.76	60.11	81.40	71.56	70.66
Listening, Oral Production	62.26	45.85	76.85	83.69	86.88	71.89	68.45
Listening, Writing	64.52	51.62	50.19	84.59	86.78	89.67	85.04
Oral Production	57.89	61.54	74.17	87.09	74.80	67.17	63.87
PhonCo, Spelling	65.21	64.92	78.77	67.52	62.93	69.21	69.09
Pronunciation, Word Recognition	77.56	73.13	64.07	55.31	65.60	72.97	71.77

PhonCo: Phonetic Coding

The results in Table 4 show that the reference composite lines tend to match one of the coordinate axes in the multidimensional **θ**-space. For example, the Cloze cluster has a reference composite line that is closest to the Dimension 1 coordinate axis – its angle with the axis is 45.69°. Also, the Listening and Oral Production cluster has a reference composite line that is closest to Dimension 2 coordinate axis, since its angle with the axis is 45.85º. The same relationship can be observed for the reference composites of the other clusters as well. Thus, each cluster defines a unique dimension corresponding to a particular language ability and aligns with a coordinate axis in the solution. Based on these results, it is clear that there exist multiple dimensions in the data that may be related to important language constructs.

3.3.7 Analysis of Most Frequently Used Test Forms

The next stage of the analysis focused on determining whether the constructs identified in the most frequently used 100 items would also appear in individual test forms. To investigate this, each of the test forms with the highest frequency of use was analyzed in the same way as the 100 most frequently used items. Because of the smaller sample size and smaller number of items, it was expected that these analyses would be less stable than the analysis of the 100 items, but the same basic pattern of results should be evident.

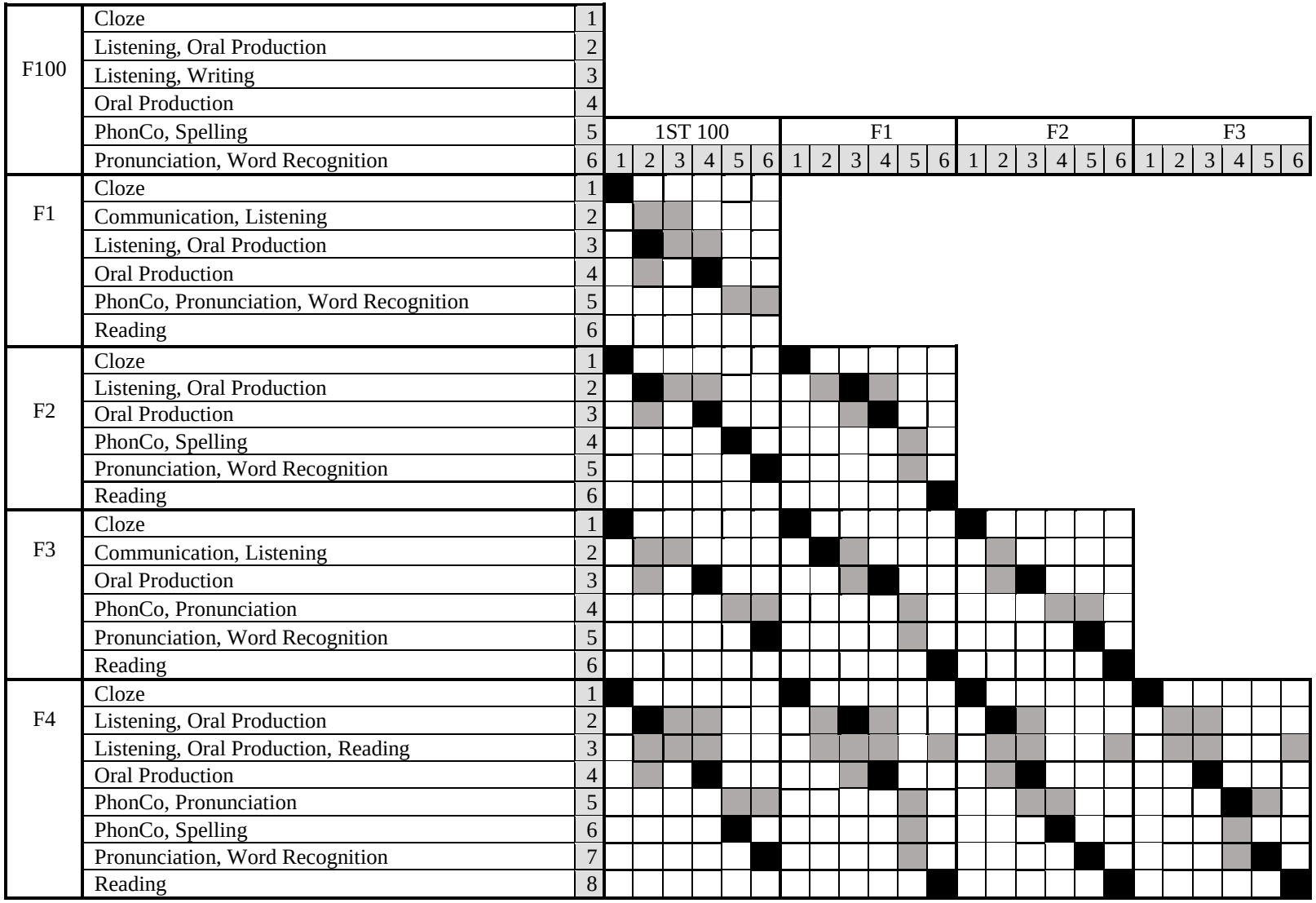
Each form was analyzed in the same way as the most frequently used 100 items – the number of dimensions was determined, a confirmatory factor analysis combined with exploratory factor analysis (CEFA) was performed using the identified number of dimensions, the angles between item pairs were computed, the items were clustered, and reference composites were determined for the clusters.

3.3.8 Consistency of Dimension Structure

To determine the consistency of structure across forms and the most frequently used 100 items, the common language constructs across four test forms were identified. Figure 2 shows a comparison between clusters identified within each form and the clusters extracted from the 100 items with highest frequencies. The number of clusters identified for the 100 items with highest frequencies (F100), F1, F2, F3, and F4 are 6, 6, 6, 6, 6, and 8, respectively. The corresponding cluster names are also alphabetically sorted in Figure 2. In the figure, the black squares indicate that clusters from the different item sets that share exactly the same language constructs whereas the grey boxes indicate only part of the constructs are the same between two clusters.

For example, Cluster 1 in F100 and F1 are Cloze, which are exactly the same, so the square is black. Clusters 2 in F100 is Listening, Oral Production while Cluster 2 in F1 is Communication, Listening. So, the square is grey because these two clusters only share the Listening construct. Also, there are no common constructs that can be identified between Cluster 2 in F100 and Cluster 1 in F1, so the cross - square is white.

Thus, F1 and F3 have very similar dimension structures. Most of the forms share some of the constructs with the 100 item set, but not all of them. That is not surprising because the 100 most frequently used items did not include all of the item types. It appears forms F2, F3, and F4 show strong multidimensional parallelism and share some of the constructs with the 100 item set.



** PhonCo: Phonetic Coding*

Figure 2: Common Clusters among Five Forms

In this research, CEFA is a combination of CFA and EFA. First, we ran cluster analysis on F100 and obtained six clusters out of form F100. Next, we selected the individual test form – the new form that could be linked back to the base coordinate system later. We sorted out the items in individual form that have exactly the same item type as those in F100. We then classified those items in individual form according to the item type and categorized them into the clusters that were already defined in F100. In other words, these items were grouped together based on the item type. These item types were clustered together according to the results from the clustering analysis of form F100. We then repeated the same procedure of CEFA for all four individual test forms. Figure 3. shows the eight clusters of Form F4 after cluster analysis using the above method.

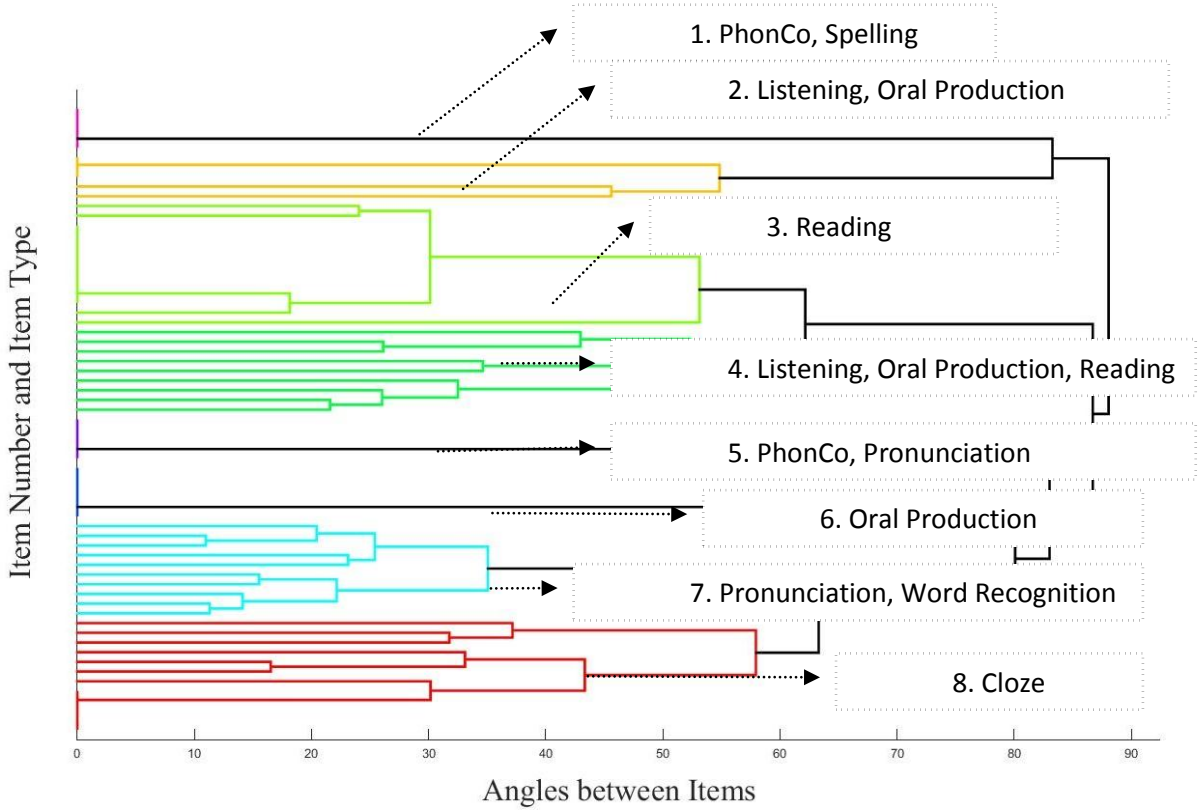


Figure 3: Clusters from Form F4

3.4 Set up the Common Coordinate System

After the application of all analyses defined above for all test forms - F100, F1, F2, F3, and F4, we need to determine the common coordinate system for rotation purpose so that the results from each individual test form can be transformed back to the base system and be compared on the common clusters.

In order to do this, we investigated the results from cluster analysis for all five forms. These results from Figure 3 show that there are some common clusters across all five forms, yet there are still some clusters that are uniquely associated with particular forms but not with other forms. Since our goal is to set up a common coordinate system so that all clusters from different individual test forms can be transformed back into this common multidimensional space, the target matrix in the nonorthogonal rotation procedure needs to be identified by the clusters in the common space. In other words, these common clusters should include all the clusters across all different test forms.

Therefore, according to the cluster analysis results for all five test forms, we take the union of all clusters identified and put the corresponding reference composites together into a big matrix composed of discrimination parameter – ***a***- parameters in a seven-dimensional space. As shown in Tables 5 to 8, there are 11 clusters in total across all five test forms. In nonorthogonal Procrustes rotation procedure, the space that contains these 11 clusters will be considered as the base coordinate system. The rotation matrix between each individual test form and this common space will be used to rotate the subscores from individual test forms into this common space.

The rotation of the coordinates will not change the location of examinee in a multidimensional space but will change the interpretation of the scores. Here, the target matrix for the nonorthogonal Procrustes rotation is defined as the identity matrix with 11 dimensions

because these 11 dimensions will align with the coordinate axes in the 11-dimensional space with the degree between each dimension equal to 90. Such orthogonal direction will help to interpret the results of scores in a more meaningful way and will also simplify the calculation procedure of nonorthogonal Procrustes rotation.

However, based on the results of cluster analysis and reference composite analysis from individual test forms, the ***a***-parameter matrix for each individual test form contains different number of dimensions and clusters. When applying the nonorthogonal Procrustes rotation, in order to solve the rotation matrix, we need to add the 0 columns to the new ***a***-parameter matrix to represent the dimensions that are not identified in that particular individual test form. Moreover, we need to add the dummy rows that represent the clusters that are not identified for that particular test form but retained in other test forms. That is, we need to investigate the ***a***-parameter matrix obtained from each individual test form and compare the clusters from that particular test form with all clusters from five test forms. By adding the dummy rows we include the clusters that are not identified for individual test form. The dummy columns indicate the number of dimensions in the base coordinate system.

Tables 5 to 8 represent the original matrix of the degrees between the coordinate axis and the reference composite for each cluster from four individual test forms, respectively. Tables 9 to 12 show the augmented matrices with additional dummy columns indicating the number of dimensions added and the dummy rows representing the clusters that are not identified for that particular individual test form.

TABLE 5:

Angles between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Six Clusters in Form F1

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7
Cloze	86.61	86.52	85.30	88.19	87.40	7.54	89.05
Communication, Listening	72.35	41.96	77.93	68.01	88.81	65.74	87.57
Listening, Oral Production	0.00	90.00	90.00	90.00	90.00	90.00	90.00
Oral Production	89.47	89.79	89.27	1.00	89.89	89.65	89.97
PhonCo, Pronunciation, Word Recognition	86.72	77.52	88.66	84.55	19.46	86.26	77.46
Reading	89.36	72.69	22.65	87.09	83.40	84.61	79.21

TABLE 6:

Angles between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Six Clusters in Form F2

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7
Cloze	86.78	85.33	87.96	87.59	89.82	6.63	88.69
Listening, Oral Production	29.55	80.84	84.98	77.49	78.83	87.81	69.35
Oral Production	76.03	82.17	78.45	22.42	88.45	89.13	80.45
PhonCo, Spelling	90.00	0.00	90.00	90.00	90.00	90.00	90.00
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	0.00	90.00	90.00
Reading	85.01	85.06	10.44	82.98	87.79	88.41	88.56

TABLE 7:

Angles between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Six Clusters in Form F3

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7
Cloze	89.66	89.67	88.65	89.19	89.34	1.85	89.47
Communication, Listening	35.84	87.08	80.99	56.14	86.58	88.33	88.27
Oral Production	90.00	90.00	90.00	0.00	90.00	90.00	90.00
PhonCo, Pronunciation	75.57	53.14	78.39	62.29	79.47	87.55	57.66
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	0.00	90.00	90.00
Reading	86.27	75.05	20.01	81.66	88.23	83.91	83.43

TABLE 8:

Angles between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Eight Clusters in Form F4

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7
Cloze	90.00	90.00	90.00	90.00	90.00	0.00	90.00
Listening, Oral Production	1.78	89.02	89.07	89.08	89.62	89.39	90.00
Listening, Oral Production, Reading	60.55	83.81	51.57	58.19	74.34	88.52	84.62
Oral Production	90.00	90.00	90.00	0.00	90.00	90.00	90.00
PhonCo, Pronunciation	84.25	59.93	82.61	73.33	69.35	77.87	46.63
PhonCo, Spelling	86.79	8.56	85.52	87.99	88.25	84.13	89.02
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	0.00	90.00	90.00
Reading	87.35	86.55	6.66	86.26	88.38	87.06	89.85

Tables 9 to 12 show the augmented matrix angles between the reference composites and the coordinate axes for the corresponding clusters in an eleven-dimensional space for each individual form. The rows represent the clusters while the columns indicate the eleven coordinates in the base space. The blue rows mean that these clusters are not from the original form but from the union across all five forms. The last four pink columns (the dummy columns when taking the cosine of 90° and 0°) represent the expanded dimensions added to the original seven dimensions for each individual form. Note, when creating these augmented matrices, two rules need to be followed. First, there should not be two identical dummy rows. Second, no zeros should appear in the same row so that the rotation matrix will not be singular.

TABLE 9:

Augmented Matrix of Angles between the Reference Composites and the Coordinate Axes in Eleven-Dimensional Space for Eleven Clusters in Form F1

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7	Axis8	Axis9	Axis10	Axis11
Cloze	86.61	86.52	85.30	88.19	87.40	7.54	89.05	90.00	90.00	90.00	90.00
Communication, Listening	72.35	41.96	77.93	68.01	88.81	65.74	87.57	90.00	90.00	90.00	90.00
Listening, Oral Production	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
Listening, Oral Production, Reading	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
Listening, Writing	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00
Oral Production	89.47	89.79	89.27	1.00	89.89	89.65	89.97	90.00	90.00	90.00	90.00
PhonCo, Pronunciation	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00
PhonCo, Pronunciation, Word Recognition	86.72	77.52	88.66	84.55	19.46	86.26	77.46	90.00	90.00	90.00	90.00
PhonCo, Spelling	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00
Reading	89.36	72.69	22.65	87.09	83.40	84.61	79.21	90.00	90.00	90.00	90.00

TABLE 10:

Augmented Matrix of Angles between the Reference Composites and the Coordinate Axes in Eleven-Dimensional Space for Eleven Clusters in Form F2

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7	Axis8	Axis9	Axis10	Axis11
Cloze	86.78	85.33	87.96	87.59	89.82	6.63	88.69	90.00	90.00	90.00	90.00
Communication, Listening	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00
Listening, Oral Production	29.55	80.84	84.98	77.49	78.83	87.81	69.35	90.00	90.00	90.00	90.00
Listening, Oral Production, Reading	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
Listening, Writing	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00
Oral Production	76.03	82.17	78.45	22.42	88.45	89.13	80.45	90.00	90.00	90.00	90.00
PhonCo, Pronunciation	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00
PhonCo, Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00
PhonCo, Spelling	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00
Reading	85.01	85.06	10.44	82.98	87.79	88.41	88.56	90.00	90.00	90.00	0.00

TABLE 11:

Augmented Matrix of Angles between the Reference Composites and the Coordinate Axes in Eleven-Dimensional Space for Eleven Clusters in Form F3

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7	Axis8	Axis9	Axis10	Axis11
Cloze	89.66	89.67	88.65	89.19	89.34	1.85	89.47	90.00	90.00	90.00	90.00
Communication, Listening	35.84	87.08	80.99	56.14	86.58	88.33	88.27	90.00	90.00	90.00	90.00
Listening, Oral Production	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
Listening, Oral Production, Reading	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00
Listening, Writing	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00
Oral Production	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
PhonCo, Pronunciation	75.57	53.14	78.39	62.29	79.47	87.55	57.66	90.00	90.00	90.00	90.00
PhonCo, Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00
PhonCo, Spelling	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00
Reading	86.27	75.05	20.01	81.66	88.23	83.91	83.43	90.00	90.00	90.00	0.00

TABLE 12:

Augmented Matrix of Angles between the Reference Composites and the Coordinate Axes in Eleven-Dimensional Space for Eleven Clusters in Form F4

Clusters	Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7	Axis8	Axis9	Axis10	Axis11
Cloze	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00
Communication, Listening	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
Listening, Oral Production	1.78	89.02	89.07	89.08	89.62	89.39	90.00	90.00	90.00	90.00	90.00
Listening, Oral Production, Reading	60.55	83.81	51.57	58.19	74.34	88.52	84.62	90.00	90.00	90.00	90.00
Listening, Writing	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00
Oral Production	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00	90.00
PhonCo, Pronunciation	84.25	59.93	82.61	73.33	69.35	77.87	46.63	90.00	90.00	90.00	90.00
PhonCo, Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00
PhonCo, Spelling	86.79	8.56	85.52	87.99	88.25	84.13	89.02	90.00	0.00	90.00	90.00
Pronunciation, Word Recognition	90.00	90.00	90.00	90.00	0.00	90.00	90.00	90.00	90.00	90.00	90.00
Reading	87.35	86.55	6.66	86.26	88.38	87.06	89.85	90.00	90.00	90.00	0.00

3.5 Two Ways of Linking Different Items on Common Multidimensional Scale

3.5.1 Common-Items Linking

First, we selected a base form. Next, within each subdomain we found common items between the base form and forms that needed to be linked to the base form. Finally, we scaled those common items onto the common scales for each subdomain.

We selected F100 as the base form, since it was extracted based on the sorted data with highest frequencies across all forms and all examinees. Forms F100 and Form 1 have 16 common items, which were distributed on six common subdomains C1 – C6. Form 1 has 65 items in total. Among those uncommon items, which are 47 items = 65 – 16, we found the items that have item types exactly the same as those in the six common constructs. In FlexMIRT, we ran EFA by fixing the item loadings to 0 so that the rest of the uncommon items could be loaded on particular dimensions corresponding with the subdomains. However, the estimation of a-parameters from FlexMIRT did not converge well. The a-parameters of these common items were not accurately calibrated. It might be due to the fact that there were few common items between the two forms. It might be also due to the reason that some items were dichotomously scored while others were polytomously scored.

Table 2 shows the common items among different forms. Table 2 shows that there are small numbers of common items between the individual forms and the base form – F100. So, the common reference composite method was considered as a better linking method for this research. Before running FlexMIRT to obtain the accurate a-parameters, some score categories needed to be collapsed so that the distribution of score category was reasonable for calibration.

3.5.2 Non-equivalent Group Common Reference Composite Linking

The idea is for each form different reference composites represent different clusters. After identifying common reference composites between the base form and the new form, we used the Procrustes rotation to rotate the common reference composites from the new form back onto the base form. Given the rotation matrix, we then rotate examinees' subscores on each reference composite from new form onto the same reference composite in the base form. Therefore, for each individual new form, there is a corresponding rotation matrix associated with the base form. Finally, when obtaining all the rotation matrices for each new form, we could start rotating all the subscores from individual forms on the identified clusters onto the base form. Thus, we could compare the subscores because they are all in the same multidimensional base space.

Moreover, the sample size for each test form varies, so in this research we applied the non-equivalent group common reference composite equating.

The reference composite is based on **a'a** matrix (the eigenvector corresponding with the largest eigenvalue from the **a'a** matrix). So, we needed to find the common clusters between each form.

In FlexMIRT, we ran CFA on those items that were classified into clusters defined in F100 according to their item types. Meanwhile, for items whose item types do not exist in F100, we ran EFA. In other words, we let those items freely load on each dimension, for we have no idea which clusters these items could be grouped together given the correlations among the data set. We ran CEFA using FlexMIRT for each individual form F1 – F4 and obtained the **a**-parameters associated with each item in seven-dimensional space according to the results from parallel analysis in F100.

There are two major reasons that we ran the CEFA to calibrate the **a**-parameters using multidimensional generalized partial credit model. First, we ran EFA directly for each individual test form, but the results from clustering were not distinguishable. That is, the items could not be distinctively classified into separate clusters. This was due to the lacking of degrees of freedom. Second, for each individual test form, by forcing to combine items that have the same item type as those in the first 100 form together, FlexMIRT could ran EFA very efficiently due to the control of degrees of freedom among the data set. This is because for individual test forms there are not enough data – examinee responses to calibrate the item parameters for the multidimensional item response models.

To sum up, our goal was to link each individual form to the base form. In other words, the base form defines the base space while the individual forms define the new space. In order to report subscores in a multidimensional space, we defined a common space where constructs from different individual forms could be rotated into the common space so that the subscores were comparable.

After running the cluster analysis for form F100 and four individual forms F1 – F4, we included all the clusters across all five forms and sorted the clusters in alphabetic order. There are 11 clusters in total. For each individual form, there was a corresponding 11-by-7 reference composite matrix with 11 clusters as rows and 7 dimensions as columns. These four matrices indicate the clusters identified in four individual coordinate systems. Each cluster is represented by the corresponding reference composite using the cosine of the degree between the reference composite and the coordinate axis. Thus, the target matrix in the nonorthogonal Procrustes rotation procedure is represented by the 11 – by – 11 identity matrix, which includes all clusters across all five forms.

When linking two forms in a multidimensional space, we rotated the new space onto the base space. We applied nonorthogonal Procrustes rotation method to obtain four different rotation matrices, respectively. One of the four rotation matrices was provided in Table 13 for demonstration purpose. The columns and rows indicate the 11 dimensions in the base space.

TABLE 13:

Rotation Matrix for Form F4

Axis1	Axis2	Axis3	Axis4	Axis5	Axis6	Axis7	Axis8	Axis9	Axis10	Axis11
-0.01	-0.02	1.01	-0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.01	-0.05	-0.80	1.68	0.00	-0.81	-0.23	0.00	0.00	-0.37	0.00
0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-0.31	-0.72	0.00	-0.31	0.00	-0.27	1.50	0.00	0.00	-0.44	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
-0.10	-0.97	0.01	-0.12	0.00	0.03	-0.01	0.00	1.00	0.01	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
-0.06	-0.01	0.75	-1.66	0.00	0.74	0.22	0.00	0.00	0.34	1.00

With the four rotation matrices, we could rotate subscores from each individual form onto the base form. After obtaining these four rotation matrices, we postmultiplied these four matrices by the estimated abilities for examinees in each individual form. Equation 31 explains the details.

$$\underline{\boldsymbol{\theta}} = \mathbf{M}^{-1}\boldsymbol{\theta}', \tag{31}$$

The rotated $\boldsymbol{\theta}$ abilities are the subscores linked back to the base form from each individual test form after multidimensional linking and equating.

For those clusters that do not belong to a particular form, the rotated subscores are not meaningful because they do not represent any constructs that the form was designed to measure. Therefore, we did not report the subscores for the clusters that were not identified in that particular individual form. When rotating subscores from new form back onto the old form, we

would get the values of subscoes for those clusters, which were not identified in that particular form. However, these values are meaningless because they just indicate the numbers that we obtained through the mathematical calculation. In other words, they are mathematically meaningful, but psychometrically meaningless.

Table 14 shows the subscores after the rotation from Form F4 to the base form. Since there are 432 examinees, only the first 10 examinees rotated subscores were provided. The highlighted blue columns correspond with the clusters that are not originally from Form 4. Thus, when interpreting the results of subscores, the subscores from these clusters will be discard because they do not represent the true language constructs measured by the items.

TABLE 14:

The Rotated Subscores after Nonorthogonal Procrustes Rotation for Form F4

Examinee ID	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
1	0.39	0.05	0.59	0.23	0.00	-0.24	-0.80	0.00	0.11	0.52	0.14
2	-0.62	-1.52	1.00	-1.60	0.00	-1.34	-1.29	0.00	-1.70	-1.16	-1.79
3	2.81	3.43	-1.26	2.92	0.00	1.20	2.32	0.00	3.94	2.33	3.81
4	2.34	2.93	-1.02	2.18	0.00	0.94	2.10	0.00	3.33	1.24	2.97
5	0.23	0.30	0.60	0.77	0.00	-0.36	0.03	0.00	0.41	0.96	0.78
6	2.64	2.53	-1.31	1.95	0.00	1.19	1.83	0.00	2.93	1.53	2.61
7	2.63	3.03	-1.80	2.51	0.00	1.47	1.83	0.00	3.48	2.15	3.50
8	2.23	1.83	-1.67	0.78	0.00	1.11	1.35	0.00	2.08	-0.81	1.86
9	0.34	-0.20	0.10	-0.18	0.00	-0.07	-0.13	0.00	-0.17	-0.68	-0.03
10	-0.09	-0.42	1.26	-0.44	0.00	-1.47	-1.08	0.00	-0.44	-0.27	-0.28

CHAPTER 4
CONCLUSIONS AND IMPLICATIONS

This research addresses major issues on how to report subscores in a multidimensional structure. First, it shows how to report subscores for a test that is well designed based on a unidimensional model. That is, it introduces a new methodology to transport the data from a unidimensional space into a multidimensional space. Almost all current operational test programs are designed using unidimensional models. Test scores or composite scores are reported on a unidimensional scale. Although researches endeavor to apply different statistical and psychometric analyses to report subscores, the multidimensional structure analysis of the data set itself is still a mysterious era that few studies have investigated.

In this research, in order to explore how to report subscores from a test that is designed using a unidimensional model to multidimensional subscales, we applied different statistical and psychometric analyses using multidimensional item response theory. Note, multidimensional item response theory is the major method we used in this research. This research does not focus on developing a statistically significant multidimensional item response model to report subscores or to compare the efficiency of reporting subscores using different multidimensional item response models. Instead, the contribution of this research is to use multidimensional item response theory to develop a scientific procedure that can be applied to report valid and reliable subscores when the test is designed using a unidimensional item response model.

In order to achieve this research goal, we applied different methods based on multidimensional item response theory. They are dimensionality analysis, reference composite analysis, and multidimensional item linking and equating. The dimensionality analysis includes the parallel analysis, exploratory factor analysis, confirmatory factor analysis, and cluster analysis. The reference composites of different clusters were calculated based on the results

from cluster analysis. The direction of the reference composite indicates the direction of the best measurement of a set of items in a particular cluster. The multidimensional linking and equating provided a plausible way to compare subscores across different individual test forms, which are designed using a unidimensional item response model.

For PTEA, the parallel analysis shows that seven dimensional space can capture the variations among the data set – F100. Therefore, setting up a multidimensional space with seven coordinate axes can be considered as a reasonable and stable coordinate system.

We originally wanted to set this seven-dimensional space as a base space in multidimensional linking and equating. However, there are not many common clusters between F100 and the four individual test forms F1 – F4. Also, there are few common items between pair of test forms among the individual test forms as shown in Table 2. Therefore, we applied the common reference composite design to link and equate test forms. Moreover, instead of using the coordinate system from F100 as the base space, we set up an 11-by-11 target identity matrix as the base space. These eleven coordinate axes represent all the clusters among all five forms from F100 to four individual test forms.

This research started with a very simple but interesting idea that is how to report subscores that can be comparable when the test is well-designed by a unidimensional model. This research idea covers the topic of test design, dimensionality analysis, scaling, linking, equating, and score reporting. It includes the development of item response theory from unidimensionalilty to multidimensionalilty. All the statistical and psychometric analyses fulfill our research goal.

The major purpose of the set of analyses reported here was to determine whether it would be meaningful to report subscores when the item response data can be well fit by a unidimensional IRT model. This was done by determining if the item response data provided evidence that multiple constructs were being assessed, and if there were, would those constructs replicate across forms? Further, do the identified constructs yield meaningful information when reported?

Moreover, the results for the dimensional analysis clearly show that, even though the overall data set is well fit by the unidimensional Rasch model, that multiple dimensions are still needed to explain the inter-relationships between the responses to test items in these data sets. The largest data set with 100 items suggests that seven dimensions are needed to represent the relationships in the data, but this data set does not include all of the item types. That suggests that more dimensions might be needed for typical test forms. Unfortunately, the sample sizes for the test forms are too small for detailed multidimensional analysis, but the pattern of results across the forms clearly indicates that multiple dimensions are needed. As more data are collected, a common structure may be identified. The analyses of the data on individual forms suggest that six to eight dimensions are needed, a result consistent with the 100 item analysis.

In conclusion, this study explores the support for the validity of the multidimensional structure across multiple test forms when the test was originally designed for a unidimensional scoring procedure using the Rasch model. Through the analysis, we can support the use of subscores for reporting. The analyses suggest that six to eight dimensions are needed to represent the constructs assessed by the different test forms.

The analysis of data set 1 - 100 items with highest frequencies across all test forms - showed the very distinct seven-dimensional solution was needed to accurately describe the relationships between the test items and the current sample of examinees. The analyses of data sets from four test forms were consistent with the 100-item analyses, supporting 6 to 8 dimensions, even though the samples were small. Figure 2 shows there was a consistency of the dimension structure across five data sets, indicating the language constructs can be replicated across multiple forms. Therefore, the subscores on the sets of items in these clusters provide meaningful differences in English skills.

4.1 Solutions to Research Questions

Research Question 1. How many distinct dimensions are needed to accurately describe the relationships between the test items for the current heterogeneous sample of examinees? In particular, is more than one dimension needed?

The analysis results show that there are seven distinct dimensions needed to capture the variation among the data sets. The first hundred items with highest frequencies indicate that a multidimensional structure with seven coordinate axes is sufficient to represent the relationships between the test items and the heterogeneous sample of examinees. Although for each individual test form the number of dimensions identified to represent the relationship among the data set is either less or larger than the seven dimensions, it is due to the small sample size of the data sets. For individual forms, the number of examinees is around 450. Compared with the frequency of 36,938 for the first hundred item data set, the results from individual forms are not as stable as those from the F100 form. That is why the number of dimensions identified from the parallel analysis for each individual form varies depending on different data sets. Also, the

clusters of items classified from each individual form are very different from each other and different from the first 100 item set, too. The research proves that even though a test can be well-designed using a unidimensional item response theory model, the data sets can still require a multidimensional structure to represent the relationships among the items and examinees.

Research Question 2. If more than one dimension is needed to represent the relationships among the test items for the current sample of examinees, are there clusters of items that are sensitive to distinct combinations of skills and knowledge and are these clusters related to known constructs of language performance?

The first research question answered that a seven-dimensional coordinate system is sufficient to represent the relationships among the data sets. The next question would be how items can be clustered together within this seven-dimensional space. It is like in a three-dimensional space, for example, in a university campus there are many departments. If we consider these departments as clusters and people on campus as items, then our questions here is how can we classify people together within this three-dimensional space? There are different ways to cluster people under this circumstance. We can group people according to their academic major. Let's say people whose research deals with psychology are usually grouped together in the building of psychology department. Therefore, according to this way classification, people on campus can be classified into different buildings.

Same philosophy applies here in clustering analysis of items in a data set. In a seven-dimensional space, how can we cluster items that are sensitive to distinct combinations of skills and knowledge? We ran the clustering analysis using the Ward's method by grouping items according to the angles between the item arrows. As proved in previous sections, the angles between item arrows indicate the correlations of the underlying constructs that these two items

can measure. The smaller the angles, the closer the items' arrows will be, and the more similar the underlying constructs and skills that the items are measuring. In this way, the items are classified into clusters that are sensitive to distinct combination of constructs and skills.

In order to identify whether these clusters of items are related to known constructs of language performance, we referred to Carroll's book which explained how each different item type in PTEA in each cluster can be related to an appropriate identified language constructs. Thus, the clustering analysis classified items using the psychometric and statistical analysis. After obtaining all these clusters, we investigated all the items regarding their item types within each cluster. Then, we applied Carroll's way of labeling these clusters using appropriate constructs related to language performance.

In other words, if Carroll's identification of these clusters is from a cognitive and linguistic perspective, then the reference composite analysis quantified these clusters by calculating the degree of angle between each reference composite line that represent each cluster and each coordinate axis. Tables 5 – 8 show these angles between each reference composite representing each cluster and the corresponding coordinate axis within each cluster.

Among the five forms in this research, F100 is the most stable one because it has the largest sample size with around 39,000 examinees. It contains the first hundred items with the highest frequency across all test forms. Therefore, the clusters in F100 did not cover all of the item types. F100 is larger sample but with more items while the four individual forms have smaller sample size but less items per content area. That is why Listening construct was divided up in F100 but not in Form F1.

Research Question 3. If meaningful clusters can be identified, are they specific to one form of the test or do similar clusters appear for more than one form? That is, do multiple forms replicate the complex structure of the language constructs?

There are five forms analyzed in total for this research. The first form – F100 is the form that contains the first hundred items that have the highest frequencies among all examinees across all 164 test forms. The second to fifth forms – F1 to F4 are individual forms that have the highest frequency around 450 among all 164 test forms. The multidimensional analysis requires data sets have sufficient responses. So, we first sorted the individual test forms according to the frequency in descending order. The first four individual test forms that have the highest frequencies around 450 were selected as the individual test forms in this research. We set up the ideal frequency number as 450 because it would return reliable results in multidimensional analysis.

In order to check for the consistency of the dimensional structure across these five forms, we ran the parallel analysis, exploratory and confirmatory factor analysis, cluster analysis, reference composite analysis for all these five forms from F100, F1 to F4, respectively. Figure 3 shows the pattern of the clusters across all five forms. Those black squares indicate there are common clusters identified across different forms. That is, these different forms replicate the complex of multidimensional structure of language constructs. The purpose for the factor analysis was to calibrate the item parameters using the multidimensional item response theory models so that we could link different test forms when report subscores.

Research Question 4. If replicable clusters can be identified in each test form, how to link and equate different test forms so that subscores from examinees taking tests on different test dates in different places can be comparable and interchangeable?

The basic is to first identify a base space and then rotate the subscores for each cluster from each individual test form back onto the base space. Since the sample size for each test form varies, in this research we applied the non-equivalent group common reference composite equating.

There are two major steps here. First, we defined the base space as an 11-by-11 identity matrix. The eleven coordinate axes represent eleven clusters across all five forms. In other words, each form has different clusters and these eleven clusters are the "union" of these different clusters so that each cluster from individual form will have a corresponding coordinate axis in the base form. Second, we applied the nonorthogonal Procrustes rotation to rotate the reference composites corresponding with each cluster in each individual test form back onto the base form. There is a one-to-one relationship between each individual test form and the base form that is quantified by a unique rotation matrix. Using that rotation matrix, we could report subscores for each cluster from individual forms onto the multidimensional scale defined in the base form. When all the subscores are rotated back onto the base form, they can be compared and used interchangeably.

4.2 Implications

This research addresses an essential question when reporting subscores for a real testing programs. Can a multidimensional structure be identified and supported as the basis for reporting subscores even though a test was originally designed for reporting a single score based on a unidimensional IRT model? The results of these analyses suggest that the answer to the question for the case analyzed here is "yes".

Three issues related to subscore reporting were addressed through this research. First, this research demonstrated statistical methods for identifying a multidimensional structure for subscore reporting on a test designed to support a unidimensional scale. The implication of this result is that even tests that result in item response data that can be well fit by a unidimensional IRT model may still tap into multiple skills and abilities. The scale that results from a unidimensional IRT analysis is a composite of those skills and abilities. If sufficient examinees and items are available, it is possible to tease out the skills and abilities that go into the composite defined by the unidimensional model.

For each cluster in an individual test form, there is a corresponding reference composite line. These reference composite lines tend to rescale the locations of examinees in a multidimensional space onto a unidimensional scale. Each reference composite represents a unidimensional scale of a subscore. When equating different test forms, the scales associated with the reference composites were rotated back onto the scales in the base form. Such transformation of subscores among different unidimensional scales in different multidimensional spaces across different individual test forms is one of the major contributions from this research.

The second issue is that large assessment programs use multiple test forms and the dimensional structure needs to be replicated over the forms taking into account of missing data. This means it is necessary to check whether the number of dimensions and the structure generalizes over multiple test forms. The scores received from multiple test forms and administrations should be comparable (Wendler & Walker, 2006, p.446). Moreover, "Validity is a unitary concept . . . [it] refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores (Messick, 1993, p.13)." Therefore, for large-scale testing programs, in order to provide

the validity evidence supporting the number and types of subscores, we need to provide evidence for a consistent dimensional structure across multiple test forms with large number of missing data.

The third issue is multidimensional linking and equating of different test forms to report subscores that can be comparable. Currently, the multiple forms of the test used in this research are equated using unidimensional IRT linking. In this study, we first confirmed the dimension structure across multiple forms. Then, we equated the individual test forms using MIRT linking procedures by setting up the base coordinate system and solving the nonorthogonal Procrustes rotation matrix for each individual test form.

Overall, this research introduces a new method to report subscores using multidimensional item response theory. It also applies a new method to investigate whether the content classification of item types is appropriate or not. In test design, the content that each item is measuring is usually defined by content experts. Yet, whether the items measure what they are supposed to measure is unclear until the data sets combined with responses from examinees are analyzed. This research provides a valid and reliable methodology for dimensionality analysis and subscore reporting.

There are several applications of the methods presented in this study. One option is they can be generalized to other large-scale testing programs with different test content that report subscores for different test forms. Also, this dimensional analysis method can be applied to identify different cognitive constructs. Finally, the dimension structure analysis can help test developers to revise the test specification. This can improve the test validity and reliability as well as the accuracy of subscore reporting for testing programs.

4.3 Limitations and Future Studies

The methods in this research demonstrated how to report subscores across different test forms using multidimensional item response theory. It started with the test form that has the highest frequency of item access. For each individual form, the analyses were analyzed based on the average sample size of 450. Such small sample sizes might cause uncertainty and unstableness when running the multidimensional analyses. Also, it might not be sufficient to calibrate the multidimensional item parameters, such as the discrimination parameters in MIRT models.

The reliability of the subscores are not provided because the subscores are IRT scores with the average standard error around 0.30. For future study, there could be research discussing the computation of reliability given the methodology of reporting subscores in this research.

The other factor that may influence the results is that there are many missing values in the total data set. For example, in the F100 form – the first hundred items with highest frequency, we used the linear regression imputation to estimate the missing values using the existing scores. Such missing-data estimation worked well for the analysis in this research. However, it is unclear whether it would work for other future research with different data sets and missing value patterns.

For future studies, it would be ideal to use more data sets without a large amount of missing values. Also, the sample size for each test form could be enlarged in order to run a valid item calibration using multidimensional item response theory models. It would reduce the random errors in statistical and psychometric analysis. The reliability of subscores would also be improved.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

Ackerman, T., & Shu, Z. (2009, April). *Using confirmatory MIRT modeling to provide diagnostic information in large scale assessment*. Paper presented at the meeting of the National Council of Measurement in Education, San Diego, CA.

American Educational Research Association (AERA), American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.) *Statistical Theories of Mental Test Scores.* Addison – Wesley, Reading, MA.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: The University of Cambridge Press.

Dorans, N.J., Pommerich, M, & Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Ferrara, S., & DeMauro, G. E. (2006). Starndardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (5th ed., pp. 579 - 621). Westport, CT: Praeger.

Haberman, S. J. (2008a). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Harman, H.H. (1976). *Modern factor analysis (3rd ed.).* Chicago, IL: The University of Chicago Press.

Harris, D. J., & Hanson, B. A. (1991, March). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Holland, P.W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (5th ed., p. 196). Westport, CT: Praeger.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 32*, 179 – 185.

Houts, C. R., & Cai, L. (2013). flexMIRT® user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Hozinger K.J., & Harman, H.H. (1941). *Factor Analysis: A Synthesis of Factorial Methods.* Chicago: The University of Chicago Press.

Kelly, T.L. (1914). Comparable measures. *Journal of Educational Psychology, 5*, 589 – 595.

Kim, J.L. (2001). *Proximity Measures and Cluster Analyses in Multidimensional Item Response Theory.* (Unpublished doctoral dissertation). Michigan State University.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.).* New York. Springer-Verlag.

Ledesma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation 12*, 1-11.

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS RB – 55 – 23). Princeton, NJ: ETS.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 60*, 523 – 547.

MATLAB and Statistics Toolbox Release 2015a, The MathWorks, Inc., Natick, Massachusetts, United States.

Messick, S. (1993). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed. pp. 13-103). Phoenix: The Oryz Press.

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

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L.Linn (Ed.), *Educational measurement* (3rd ed., pp. 335 - 366). New York: Macmillan.

Monaghan, W. (2006). *The facts about subscores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved December 12, 2013, 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*, 159–176.

Muthén, L. K., & Muthén, B. O. (2005). Mplus User's Guide. (Version 3). Los Angeles, CA: Muthén & Muthén.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: The National Academies Press.

Pearson Longman (2010). *The Official Guide to PTE: Pearson Test of English Academic.* Hong Kong: Pearson Longman Asia ELT.

Pinter, R. (1914). A comparison of the Ayres and Thorndike handwriting scales. *Journal of Educational Psychology, 5,* 525 – 536.

Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* pp. 194 – 204.

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

Reckase, M.D. (2009). Analyzing the structure of test data. In *Multidimensional item response theory* (pp. 179-231). New York, NY: Springer.

Reckase, M.D. (2009). Transforming parameter estimates to a specified coordinate system. In *Multidimensional item response theory* (pp. 233-273). New York, NY: Springer.

Reckase, M.D. (2009). Statistical descriptions of item and test functioning. In *Multidimensional item response theory* (pp. 113-135). New York, NY: Springer.

Reckase, M.D. (2009). *Multidimensional item response theory.* New York: Springer.

Reckase, D.M., & Xu, J.R. (2015). The evidence for a subscore structure in a test of English language competency for English language learners. *Educational and Psychological Measurement, 75*(5), 805 – 825.

Reise S.P., Waller, N.G., and Comrey A.L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* pp. 113 – 135.

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

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150-174.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26* (4), 21–28.

Sinharay, S., Puhan, G. , & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30* (3), 29–40.

Starch, D. (1913). The measurement of handwriting. *Journal of Educational Psychology, 4,* 445 – 464.

Stone, C.A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63-86.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Thurstone, L.L. (1947). *Multiple-factor Analysis: A Development and Expansion of The Vectors of Mind.* Chicago: The University of Chicago Press.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*,*61*, 287–307.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A. & Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Erlbaum Associates.

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

Weiss, A. P. (1914). A modified slide rule and the index method of individual measurements. *Journal of Educational Psychology, 5,* 511 – 524.

Wendler, C. L, & Walker, M.E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.445 - 467). Mahwah, NJ: Erlbaum Associates.

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

Yen, W.M, & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (5th ed., pp. 111 - 153). Westport, CT: Praeger.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.