This is to certify that the
dissertation entitled

MULTIDIMENSIONALITY AND ITEM PARAMETER DRIFT:
AN INVESTIGATION OF LINKING ITEMS
IN A LARGE-SCALE CERTIFICATION TEST

presented by

XIN LI

has been accepted towards fulfillment
of the requirements for the

Ph. D. degree in Measurement and Quantitative Methods

Major Professor's Signature

9/19/08

Date

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
0 3 0 8 10		

MULTIDIMENSIONALITY AND ITEM PARAMETER DRIFT: AN INVESTIGATION OF LINKING ITEMS IN A LARGE-SCALE CERTIFICATION TEST

By

Xin Li

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2008

ABSTRACT

MULTIDIMENSIONALITY AND ITEM PARAMETER DRIFT: AN INVESTIGATION OF LINKING ITEMS IN A LARGE-SCALE CERTIFICATION TEST

By

Xin Li

Common items are widely used to equate test scores of multiple forms or to link test scores at different grade levels. Violations of the assumptions of IRT models may distort the parameter invariance requirement for the use of linking items for repeated use over time. Unidimensionality is one of the primary assumptions. However, linking items are meant to be representative of the whole test and are likely to be sensitive to multiple dimensions if that is the intent of the test. Parameter drift occurs if the statistical properties of items change over time (Goldstein, 1983).

In this study, a large-scale certification test dataset was used that included 30 linking items that were administered three times within six years. Simulated responses for these common items were generated from the "true" item parameters from a concurrent calibrated using the data from three years of administration. In addition, samples were randomly selected from the real data. Item parameters were estimated and linked to the same scale using both unidimensional and multidimensional techniques. Compared to the simulated data under the assumption of no drift in item parameters, the property of parameter invariance for the real data samples was examined for different latent trait distributions at different time points.

The results confirmed that the potential effect of multidimensionality was associated with the item parameter drift (IPD) for the same set of common items using four different dimensional compositions. Also discussed were limitations for the artificially constructed compositions of actual item response and robustness of IPD detection. Future areas of research are suggested.

Dedicated to my beloved: Guanghui Liu and Lucy Zixi Liu

ACKNOWLEDGMENTS

I would like to express my earnest gratitude to my advisor Dr. Mark Reckase. Dr. Reckase has always been an admirable and brilliant mentor to me, has shown substantial support for my study and work with his inspiration and encouragement, and has strengthened my determination in pursuit of my dream. His persistent efforts and patience have assisted me through the most challenging but rewarding experiences to become a professional researcher.

I also wish to show sincere appreciation to members of my guidance committee, Dr. Yeow Meng Thum, Dr. Kimberly Maier, and Dr. James Stapleton for their constructive comments and insightful suggestions on my work. I have been extensively benefited from their marvelous instruction and treasured their friendship during my entire doctoral study and particularly in writing and revision of this dissertation.

This dissertation research was supported by grants from the Spaan Fellowship offered by the University of Michigan English Language Institute (ELI-UM). I am grateful for their permission to use the data and offering funding for this study. I would like to thank Dr. Amy Yamashiro, Dr. Jeff Johnson, Eric Lagergren, and other colleagues at ELI for companionship and collaboration.

I owe countless thanks to my husband, Guanghui Liu, my parents, my parents-in-law, and my friends for years of unconditional love and support. Their encouragement and belief have pushed me to finish this work.

TABLE OF CONTENTS

CHAPTER 5

LIST OF TABLES

LIST OF FIGURES

(Images in this dissertation are presented in color)

CHAPTER 1

INTRODUCTION

1.1 Research Background

Large-scale testing programs often require comparisons of test scores obtained from multiple forms of the same test or even from tests of different difficulty levels. The purpose is to maintain test security over time or to measure changes in test performance without repeating identical questions. Items from multiple forms or different sets are considered as separate subsets of a large pool selected to measure the same construct. However, the extent to which scores from these different sets of items can be used interchangeably is determined by how the test scores are linked. Linking refers to the general process of transforming scores from one test to those from another to obtain equivalent trait scores (Holland & Dorans, 2006; Kaskowitz & Ayala, 2001; Cook & Eignor, 1991; Dorans, 1990). Comparable score scales are constructed by placing test scores from different forms on the same metric. Linking functions are estimated to describe the relationship between scores from two or more tests.

A collection of methods for estimating test linkages have been developed to make the outcome scores comparable in most practical testing circumstances. Different methods for estimating test linkages require a variety of criteria and statistical procedures. As defined in *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), equating refers to the process of placing scores from different test versions on a common scale on the condition that these distinct forms are equivalent. Five stringent conditions are assumed for equivalent test forms including equal constructs, equal reliability, symmetry of test structures, equity for the assessments and invariance of examinee populations to be linked (Dorans & Holland, 2000; Dorans, 2004, 2000). If these conditions are met, the test scores from parallel forms can be converted to the same reference scale and used interchangeably. The score conversion is theoretically constrained for use with alternate forms of the same test. Two types of test equating are typical in practice. Given tests measuring identical content areas, horizontal equating is defined as a method of converting scores on tests at similar difficulty levels and vertical scaling is defined as a method of converting scores on tests of different difficulty levels. The purpose of horizontal equating is to allow comparisons of different groups of examinees within the same grade level using multiple forms of the same tests. Vertical scaling allows comparison of test scores of students from a variety of grades for tracking student growth across grades (Kolen & Brennan 2004, 1995).

Score conversions are also useful for scaling scores from some tests that cannot be equated. For example, scores may be aligned on a reference scale for tests of similar content areas but differing in length, languages, or format (AERA, APA, NCME, 1999). Dorans (2004, 2000) discussed another two types of test linkages that require less equivalency. Concordance describes the link between tests built to different specifications. An example is linking the ACT scores to the Scholastic Achievement Test (SAT) by a concordance function to align cut-scores for college and university admission. Prediction, however, links tests that measure different constructs. For instance, student scores on previous ACT or SAT tests are used to estimate the future grade point average expected for students in colleges or universities.

As is the case for most areas of scientific research, test equating or linking can be accomplished though the use of a variety of designs for collecting data. Single group, random groups, equivalent groups, and common-item nonequivalent groups are commonly used designs in test equating (Kolen & Brennan, 1995). The single-group design uses the same group of examinees for both tests to directly control for difference in performance. The examinees' scores on both forms are assumed independent of the order in which the forms are administered. In reality, the assumptions of this design are not likely met due to the impact of practice and fatigue. The random-group design allows a single group randomly divided into half so that each subgroup takes both tests in counterbalanced order. Due to prolonged testing time, this is rarely employed in practice. A simple solution is to administer each test to two random samples that are representatives from the same population, known as equivalent-group design. However, it is difficult to get exactly equivalent groups or administer both forms at the same time in many testing programs. The common-item design allows two samples at different administrations and improves the flexibility since only one test form is administered to each group.

As a result, it has been a common practice to insert a set of items into operational tests for repeated use across years or over multiple administrations in large-scale assessments. These common items are also referred to as anchor or linking items. They have been mostly used for equating test scores of alternate forms, scaling parameter estimates to a calibrated item pool, or linking current tests to previous versions that measure similar constructs but differ in length or format (Yu & Popp, 2005; Kolen & Brennan, 1995). Based on the examinees' performance on these common items, test scores can be placed on the same scale and become comparable for different groups of examinees. It is possible to make interpretations of these scores and to verify that the interpretations are meaningful. It is also important to check the assumptions that the properties of these linking items are invariant for multiple administrations across forms or over time.

Classical test theory (CTT) has been conventionally used for modeling the observed item and test scores. Conventional item statistics consist of item difficulty represented as proportion of correct response and item discrimination as point biserial correlations. However, both are dependent on the examinee samples and examinee scores are dependent on the test items administered (Hambleton & Jones, 1993; Lord, 1980). Only when representative samples are carefully selected, reliable item and test statistics can be used to generate parallel forms measuring the same construct. In addition, CTT is limited to test level studies due to the fact that its underlying framework assumes test scores as a combination of true scores and error scores (Hambleton & Jones, 1993). The focus is determining the error of measurement based on the total test scores from a set of items, which restricts its applications to item level analyses and many measurement situations in practice. For instance, statistics are obtained from responses to the same set of anchor items may depend on samples of examinees who take the test or the set of items being selected, which poses theoretical and practical difficulties in using CTT models for test equating (Fan & Ping, 1999).

As an alternative, models underlying item response theory (IRT) are based on the framework that examinee abilities are associated with their performances on individual items. Compared to the sample dependency of item parameters in CTT, IRT has the property of invariance for item and ability parameters given the model fit to the test data of interest (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). A linear relationship is assumed between ability estimates obtained from different sets of items and item parameter estimates obtained in different groups of examinees except for measurement errors.

The feature of invariant item parameters are primarily desirable in maintaining the item pool and linking test scores on alternate forms, which makes IRT widely used for a variety of purposes, such as test equating, score scaling, and computerized adaptive testing. Take test equating for example, score scale conversions are derived from the responses to common items under the assumption that their parameters remain the same. That is to say, characteristics of linking items are independent of diverse ability distributions of the examinees over multiple administrations. Given unchanged item parameters, the observed difference in scaled scores can be attributed to the difference in abilities across groups or measurement of growth over time.

However, the validity of scaled test scores are challenged if anchor items do not function identically over repeated administrations for the target population. That is, the statistical properties of the same items change over time (e.g., item difficulty value and item discrimination power), which is referred to as item parameter drift (IPD) (Goldstein, 1983). In other words, the estimates of item parameters are statistically different when identical items are administered in alternate test forms at different times. Of concern is the invariance of item parameters over time. Another situation is that item functioning in the same test may interact with irrelevant sample characteristics of examinee, such as gender, ethnicity, and social economic status. This is known as differential item functioning (DIF). The focus is on differences in item functioning between reference and focal groups.

Even though anchor items are in general carefully selected and secured as high-quality items, drift is likely to occur in maintaining an item pool over time or repeatedly administering these common items for linking purposes. Such effects may be expected especially for these linking items due to frequent item exposure, increasing practice effect, or inappropriate test-wise training for identical items or items that are literally similar. Items may also perform differently across years because of changes in the construct or content. Take language assessment for example, anchor items become relatively easier or less discriminating due to growing popularity of certain words and phrases or over exposure to the target population. As a result, the item may not be sensitive to the variation in knowing these terms as when they were administered previously. In particular, changes related to national, ethnic, and cultural issues may also confound estimates of item parameters.

1.2 Research Objectives and Questions

It is often assumed that applications of IRT models requires unidimensional test data indicating the variation in test performance can only be attributed to a single construct. The assumption of unidimensionality can also be met if each item in a group of items measures the same composite of traits (Reckase, Ackerman, & Carlson, 1988). Most test items in practice, however, may not meet the unidimensionality assumption but measure a combination of skills rather than one. Also, the composites of abilities may not remain the same after repeated administrations. Lack of parameter invariance suggests that the assumption of the IRT model is not satisfied and suggests misfit of the model to the data (Oshima, Raju, & Flowers, 1997; Way, Garey, & Golub-smith, 1992).

Multidimensional IRT models should be a better option and fit the data better in these circumstances. However, it remains unclear how much this theoretical explanation affects practice, why a few items exhibit variation and what is the consequence of using these items in computation of the equating or linking function.

A collection of simulation studies has investigated the impact of variation in item parameter estimates. The items identified as variant were directly related to the difference in trait correlation (Oshima & Miller, 1990) and resulting item characteristic curves and true scores (Oshima, et. al, 1997). Others found the ability estimates and pass/fail status were robust given the undetected variation in item parameter estimates (Witt, Stahl, & Bergstrom, 2003). The causes of lack of item parameter invariance have been attributed to instruction (Cook, Eignor, & Taft, 1988), context effects (Eignor, 1985), location in the test forms (Sykes & Ito, 1993), and overall characteristics of the test (Chan, Drasgow, & Sawin, 1999).

Nevertheless, most of these studies assumed unidimensionality of the test data while a few applied unidimensional models to simulated multidimensional data. Little study has done on IPD using multidimensional models. In addition, the drift of item parameters in previous researches was arbitrarily simulated but might not appropriately reflect the real changes in practice. The values representing the drift could be too small to be identified given the statistical techniques that were used. Alternatively, the specified changes in item parameters could be too substantial to be expected in real test administrations. There was a lack of empirical research about the potential consequences of test data sensitive to multiple dimensions on the invariance property of IRT parameter estimates.

This study investigated the test dimensionality of a large-scale certificatation testing program and checked the parameter invariance of linking items. Based on the "true" item parameters calibrated from the real data, simulated item responses for these common items were generated to replicate the actual distribution of examinees performance on the test. Samples were also randomly selected from each year of test administration. Item parameters for both simulated and real data were estimated and linked to the same scale using both unidimensional and multidimensional techniques. Compared to the simulated data assuming no drift in item parameters, the property of parameter invariance for the real data samples were examined for different latent trait distributions at different time points.

The primary purpose was to compare the variations in item parameter estimates obtained from the real samples to those from simulated data across three administrations using four types of dimensional structures. Both unidimensional and multidimensional techniques were applied to explore the potential impact of multidimensionality upon the degree of invariance of item parameter estimates. A variety of test structures was compared by grouping items with regard to the hypothetical framework. The equivalence of measurement models was also compared across years to verify that the construct the test was designed to measure remained invariant over time. Different combinations of sections were analyzed for both simulated and real test data to identify the impact of multidimensionality on IPD detection. To be specific, the research questions were:

1. Did the parameter estimates (item difficulty values and item discrimination levels) of the anchor items differ over cycles after the item parameter estimates for samples selected for each single year were linked

to “true” item parameters used for simulation bases on a combined sample of all three years?

2. Were the distributions of scaled item parameter estimates obtained from the simulation samples using “true” item parameters similar to those obtained from randomly selected samples of real test data?
3. Were the patterns of drift across years of administrations similar for the under the assumptions of the four different dimensional structures? To what extent was the drift of IRT item parameter estimates influenced by the violation of unidimensionality assumption in IRT models?
4. Were the differences in item parameter estimates statistically significant? If so, did item parameter drift substantially deteriorate examinees’ true scores estimates?

CHAPTER 2

LITERATURE REVIEW

2.1 Item Response Theory

The primary interest of a test is to locate the individual along the continua of underlying dimensions based on their performance on a set of items designed to measure the latent traits. Item response theory (IRT) is a theoretical framework using nonlinear mathematical models to represent the interaction between examinee performance on each item and abilities measured by the items in the test. To be specific, it expresses the probability of the correct response to an item as a nonlinear function of the latent trait measured and the parameters characterizing the item (Lord, 1980).

Assuming that the items measure the same trait that is not directly observable, IRT introduces statistical procedures to model a single dimension of an underlying construct on which examinees rely for correctly answering test items. The latent trait, denoted as θ, is estimated as a continuous unidimensional variable that explains the covariance among item responses (Steinberg & Thissen, 1995). The location of the item along the dimension (b) represents the item difficulty. Another parameter is item discrimination (a), indicating the strength to which item responses vary between people *with* θ level above or below the item difficulty. Lower asymptote (c), also known as pseudo-chance level, is the probability that a person without any knowledge for this particular item correctly answers the item.

For item *i*, the probability that examinee *j* with given ability level θ_j correctly answers the item ($X_{ij} = 1$) is assumed to be represented by the logistic function with all three item parameters (a_i, b_i, c_i) as follows:

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{\exp(1.702 a_i(\theta_j - b_i)}{1 + \exp(1.702 a_i(\theta_j - b_i)} \qquad (2.\ 1)$$

The three-parameter logistic model (Lord, 1980) was developed to apply item response theory to multiple-choice items that may elicit guessing. Other IRT models included the two-parameter logistic model (Birnbaum, 1968) with guessing parameter assumed or constrained to be zero, and the one-parameter logistic model that was originally introduced as the Rasch model (Rasch, 1960, 1961) assuming that item discriminations of all items are equal or fixed. Alternative models similar to logistic models were normal ogive models that were based on the cumulative normal probability distribution (Lord & Novick, 1968). A constant of 1.702 was used for scaling the logistic functions. The rescaled logistic function differed by less than 0.01 in probability for all values of θ compared to the normal ogive function (Birbaum, 1968). The formulation based on the logistic model was computationally simpler and has been widely used for item calibration and parameter estimation.

Detailed account of IRT models for dichotomous multiple-choice items are given in Lord (1980), Hambleton and Swaminathan (1985), Hulin, Drasgow and Parsons (1983), and van der Linden and Hambleton (1997). Other models have been developed for polytomous data and multidimensional data.

2.1.1 Features of Parameter Invariance

IRT models are statistical functions of a person's ability and the essential characteristics of the items she/he encounters in a test. Compared to the constraint of sample dependence in CTT models, invariant item and person statistics make IRT models theoretically and practically valued for measurement problems given that the model fits the data (Lord, 1980; Fan & Ping, 1999; Rupp & Zumbo, 2006).

One property of invariance is *item parameter invariance*, namely, the idea that the item statistics are independent of groups of examinees selected from the population for whom the test intends to measure. Item parameter estimates including item difficulty, item discrimination, and lower asymptote are assumed the same regardless of the distribution of ability except for measurement error (Hambleton, Swaminathan, & Rogers, 1991). The probability of correct response to any item by examinees of a given θ level depends on θ only, not by the frequency of examinees at that θ level or those at any other θ level (Lord, 1980).

The other property of invariance is *person parameter invariance*, indicating that the scaling of the latent trait does not depend on any set of items. No matter what items are administrated to him or her, an individual's true location on the latent trait continuum remains the same, which in turn determines his or her responses to different sets of items. IRT calibrates θ scores for an individual not simply based on the number of correct responses but also takes into account the patterns of correct and incorrect answer to items and the characteristics of the items. As a result, responses to any set of items can be used to estimate an individual's θ score (Weiss & Yoes, 1991; Hambleton, Zaal, & Pieters, 1991).

The test-and group- invariant estimates of examinee abilities within IRT provide an essential advantage over CTT and lead to its applications to practical issues related to test construction, item selection, score equating, and computerized adaptive testing (Stenbeck, 1992). However, the invariance property should not be constrained only to replication of the data given the model fit to the data. It is of crucial importance to assure that the test measures the same attributes in the same way in different subpopulations, known as measurement invariance. This basic requirement of the measurement procedure indicates that the relationships between test items are invariant across ability scales of examinees (Meredith, 1993; Stenbeck, 1992; Mellenbergh, 1989).

The invariance property is of fundamental importance to justify the application of IRT models. Suppose that the test domains and the target examinee population remain the same, the IRT item and ability parameters are invariant if the assumptions are met and correct models are applied. This suggests a theoretically ideal state with perfect model fit to the test data (Rupp & Zumbo, 2006; Hambleton et al., 1991). The variations of parameters may suggest the misfit of the model to the test data or that the assumption of unidimensionality is not satisfied. However, it may also be due to the fact that the test domains and/or the candidate's population have changed. In such cases, the measurement model is not the same and unidimensional IRT models should not be applied.

Though the model fits the data and the measurement model is invariant, the true parameters are typically unknown in practice and have to be estimated from examinees' responses to the items. The metric obtained from item parameter estimates is unique up to a linear transformation (Lord, 1982). The response function stays the same provided that parameter estimates fit the following linear transformation:

$$\theta_j^* = A\theta_j + B \tag{2.2}$$

$$a_i^* = \frac{1}{A}a_i, \; b_i^* = Ab_i + B, \text{ and } c_i^* = c_i, \tag{2.3}$$

where

A and B are linking coefficients representing slope and intercept of the function,

θ_j is the ability estimate for examinee *j* on the original metric,

θ_j^* is the ability estimate for examinee *j* on the target metric,

a_i, b_i, c_i are the discrimination , location and lower asymptote estimates respectively for item *i* on the original metric,

a_i^*, b_i^*, c_i^* are the discrimination, location and lower asymptote estimates respectively for item *i* on the target metric.

Of concern is the indeterminacy of IRT parameter scales since all parameter estimates within a linear transformation have identical probability of correct responses. Arbitrary constraints are typically imposed for parameter calibration. By convention, the choice of ability distribution is typically set with a mean of zero and a variance of one. Provided that the origin and unit for measuring ability are specified, the item parameter estimates can be identified with a set of values. Otherwise, there are infinite solutions to the item parameter estimates that lead to the same probability of correct item response on the condition that they fit a linear transformation.

In practice, the computation of the linking coefficients for the linking and equating method depended heavily on the accuracy and consistency of item parameter estimates (Kaskowitz & Ayala, 2001; Way, Carey, & Golub-Smith, 1992). Rupp and Zumbo (2006) conducted a comprehensive investigation of violation of the linear transformation using analytical, numerical, and visual tools. The results indicated that IRT model inferences about examinees were relatively robust toward moderate amounts of lack of invariance across a wide range of theoretical conditions. The invariance property of item parameters was theoretically attributed to meeting the assumption of the IRT model. However, in practice, the reasons for the variations in item parameter estimates were not that simple. It might be due to the heterogeneous population being studied or a complex mixture of traits being elicited by test items. It remained unclear how the theoretical relationship was associated with applications of the IRT model in common item equating and linking.

2.1.2 Assumption of Unidimensionality

IRT models specify the proportion of correct response by an examinee as a function of the examinee's ability and the characteristic curve of items. For appropriate application of IRT models, three primary assumptions are generally required about the test data: unidimensionality, local independence, and monotonicity. Unidimensionality assumes that only one latent trait is measured. Local independence requires that responses to any pair of items are statistically independent. Monotonicity indicates that the probability of correct responses increases with increases in ability.

Lord (1980) stated that "The invariance of item parameters across groups is one of the most important characteristics of item response theory. ... But our basic assumption is that the test items have only one dimension in common" (p.35). Unidimensionality is a common assumption underlying IRT models that a single construct is measured by a set of items in a test. Examinee performance is attributed to the latent trait (ability) in one-dimensional space (Lord & Novick, 1968; Reckase, 1979). The assumption of unidimensionality is also closely related to the other assumptions. Local independence is a sufficient condition for achieving unidimensionality and both concepts are equivalent. Other cognitive and affective factors that are unrelated to the construct the test designed to measure may be confounded with test performance, such as speediness, motivation, and test anxiety, which account for construct-irrelevant variations (Weiss & Yoes, 1991; Hambleton, Swaminathan, & Rogers, 1991).

Unidimensionality assumes that a single common trait is what items are designed to measure. The extent to which the assumption can be adequately satisfied in practice has been extensively researched. A great number of decision criteria have also been developed for operational purposes. A traditional method for assessing unidimensionality is to verify the presence of a dominant component or factor measured by the test (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). Factor analytic techniques have been widely used to determine the dominant factors. The linear relationship between pairs of items are summarized using a matrix of phi coefficients or tetrachoric correlations. Eigenvalues are extracted from the correlation matrix. Factors are significant if they have eigenvalues greater than one (Kaiser, 1960), account for at least 10% of the total variance (Hatcher, 1994) and exhibit a significant drop in eigenvalue plots (Reckase, 1979).

Other methods include the coefficient alpha for assessing internal consistency and the ratio of eigenvalues. Hattie (1985) proposed an index of relative changes in eigenvalues for consecutive factors to determine the number of dominant factors. The ratio of difference in consecutive factors, denoted as Factor Difference Ratio Index (FDRI), should be above three to represent a substantial contribution of the first factor.

Previously proposed methods, labeled as "traditional methods", involve linear factor analysis for assessing *strict dimensionality* that only one latent trait is required to produce monotonic and locally independent IRT models (Tate, 2003; Hattie, 1984, 1985). Stout (1987, 1990) introduced the new concept of *essential unidimensionality* as the presence of exactly one dominant dimension, which was based on the theory of essential independence. Essential independence is a comparatively weaker assumption than local independence. The item pair conditional covariance for each fixed latent trait is required to be small in magnitude and decreasing as the test length increases. A *t*-statistic was proposed to detect departures from essential dimensionality of a test data set using a computer program called DIMTEST (Stout, Nandakumar, Junker, Chang, & Steidinger, 1991). The algorithm tests whether the dimension structure of the selected set of Assessment Test items (called AT items) is homogeneous and significantly distinctive from the remaining set of items in the Partitioning Test (called PT items).

These procedures result in useful applications for finding the dominant component or factor. Statistical tests also tell whether the dominant factor is significantly distinctive from the other negligible ones. Nevertheless, it remains a question how many dimensions should be retained if more than one factor is important.

2.2 Multidimensionality

It has long been argued that most tests are inherently multidimensional suggesting test items measure more than one construct (Ackerman, Griel, & Walker, 2003; Ackerman, 1994; Hambleton & Swaminathan, 1985; Reckase, 1979, 1985; Stout, 1987). The constructs are the latent attributes that a set of test items are designed to assess and the dimensions are the number of coordinates defined in the multi-dimensional space to discriminate the sample of people studied (Reckase, 2006). Within the framework of item response theory, dimensionality is not a characteristic of items but reflect the interaction between test items and examinee capabilities.

The dimension/category framework introduced by Boeck, Wilson, & Acton (2005) describes latent structure behind manifest categories as underlying quantities depending on the person and the data via functions of probabilities of responses. Given persons within the same manifest categories, the heterogeneity in their performance at latent level suggested an internal structure. That is, different persons had different locations along the continuum of latent dimension. Otherwise, the dimensions fell into one category when everyone was located at the same latent level. In addition, the quantitative differences between manifest categories were expected to distinguish persons on the same latent dimension across different manifest categories while qualitative differences representing latent level profiles differ from one to another.

Wainer and Thissen (1996) have proposed two sources of multidimensionality for interpretation. Individual differences in examinee performance are attributable to the internal structure or characteristics of the assessments. If the factor structures of a test reflect the number of dimensions the test is designed to measure, the multidimensionality is considered as fixed for identifying domains being assessed. On the other hand, random multidimensionality represents the presence of minor dimensions that are distinctive from the target dimensions. These nuisance factors are not generated by design but reflect trivial item characteristics and lead to construct-irrelevant variance.

Reckase (1994) argued that the purpose of the assessment also affected the determination of dimensionality. If it is important to identify domains being measured and the relationships between the domains, all possible dimensions should be included and overestimation of the dimensionality is desirable. However, if the goal is to obtain ability score estimates, "the dimensionality of interest is the minimum dimensionality that provides the greatest information provided by the item responses" (p.90). In this case, overestimating dimensionality leads to an increase in numbers of parameters to be estimated, which results in extra estimation error.

Examination of a test's internal structure provides evidence for hypothesized multidimensional test blueprints (Martineau, et al., 2006) and test score interpretations (Stone & Yeh, 2006). Independent clusters serve to determine dimensionality and establish the matrix of pattern for identification purposes and trait interpretation (McDonald, 2000). Two types of dimensional structures are in general categorized for multidimensional test data. One is known as simple structure. For a set of items that are sensitive to multiple dimensions, each item is represented by only one of those dimensions (Walker, Razia, & Thomas, 2006; McDonald, 1999). That is to say, the loadings for each item are high on one particular factor but zero or close to zero for other factors. All items lie along the coordinate axes (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). Traditional dimensionality analyses focus on the simple structure while the real test data may be much more complicated. The other type of structure is a complex structure that is more common in practice where each item loads on every factor instead of being dominated by only one factor. Items lie along composites in multiple dimensional spaces (Walker et al., 2006; McDonald, 1999; Stout et al., 1996).

A clear identification of dimensionality should be substantively supported by expert judgment (Ackerman et al., 2003). Subsets of items are arbitrarily assigned in terms of test specification or the discrimination values are determined on each dimension for each item. Three approaches were advocated by Ackerman et al. (2003) as guidelines for assessing substantive dimensionality. Firstly, test specification provides the outlines of the achievement domains. A subsequent content analysis is conducted by professional experts to identify dimensions based on item content. Finally, psychological analysis determines the cognitive skills for successful response to each item.

The substantive dimensionality, however, may not be consistent with the results of statistical analyses. Walker et al. (2006) pointed out the difference between substantive and statistical dimensionality. Even though a set of test items is designed to measure substantively more than one dimension, the item response data show little variation for the other dimensions. The other factors become statistically insignificant and the test data is indeed one-dimensional. As an alternative, although most tests in use are developed to match test specifications for assessing one construct, the assumption of unidimensionality is violated due to the multidimensional nature of test items and test purposes in educational and psychological tests. Some tests, such as spelling, involve a basic skill and may fit a one-dimensional model reasonably well. Other tests, such as mathematics, may require a combination of arithmetic knowledge and reading comprehension to answer items correctly. Language tests in particular consist of subtests assessing a variety of skills for students from diverse educational, linguistic, and cultural background (Henning, Hudsan, & Turner, 1985).

A voluminous body of research literature has been accumulated for checking whether the assumption of unidimensionality is strictly or essentially met. Once the assumption is violated, another important concern is to identify the number and the nature of the latent traits that determine the variation in examinee performance on the test. Researchers have extensively studied methods for examining the internal structure of test data and providing evidence for hypothesized multidimensionality. A comprehensive review of statistical procedures for assessing test structure are available in Tate (2002, 2003), Hambleton et al. (1991), and Hattie (1984, 1985).

Both parametric and non-parametric techniques have been developed for assessing dimensionality. Most of the parametric methods assume multidimensional item response theory (MIRT) models. Many procedures have been applied to the estimation of both item and ability parameters for the models. Computer programs have also been developed for implementing these procedures. Two current programs in wide use are TESTFACT (Wilson, Wood, Gibbons, Schilling, Muraki, & Bock, 2003) and NOHARM (Normal-Ogive Harmonic Analysis Robust Method) (Fraser, 1993). The full information item factor analysis (Bock, Gibbons, & Muraki, 2003) implemented in TESTFACT uses marginal maximum likelihood to estimate parameters based on all information offered in test responses rather than a matrix of item covariance or correlation. Alternatively, NOHARM performs nonlinear item factor analysis by fitting a polynomial approximation to the multidimensional normal ogive model. Least squares estimation of item parameters are obtained from the matrix of raw product moments. According to McDonald (1981), latent trait dimensionality should be based on misfit of the models to test data. Residual analysis is a good measure for assessing model fit. A comprehensive review by Hattie (1984, 1985) also reported that residual analysis is the most effective measure out of 87 indices as decision criteria in psychometric literature for assessing unidimensionality.

Another type of research focuses on nonparametric IRT (NIRT) methods to account for the relationship between dichotomous item score without an underlying assumption that the data are drawn from a given probability distribution. The two NIRT methods include DETECT (Kim, 1994; Zhang & Stout, 1999a, 1999b), and HCA/CCPROX (Roussos, Stout, & Marden, 1998) and have been widely used to determine the number of dimensions of the item response data. In addition, DIMTEST (Stout, Froelich, & Gao, 2001; Nandakumar & Stout, 1993; Stout, Douglas, Junker, & Roussos, 1993; Stout, Nandakumar, Junker, Chang, & Steidinger, 1991) was introduced for testing hypotheses that the dimensionality of item response data is significantly different from one. A comparative study conducted by van Abswoude, van der Ark, & Sijtsma (2004) recommended methods based on covariances conditional on the latent trait, such as HCA/CCPROX, for retrieving simulated dimensionality structure and for understanding the process of clustering.

An approach, known as Parallel analysis, uses Monte Carlo simulations to determine the factors that are important (Ledesma & Valero-Mora, 2007; Hambleton, et al., 1991; Drasgow & Lissak, 1983; Horn, 1965). Datasets are generated as random samples to simulate observed item response data with equal sample size and number of variables. Eigenvalues obtained from the random data with zero correlation are plotted with those computed from the observed item responses. If the test data is unidimensional, both plots of eigenvalues should be close to each other but only the first eigenvalue extracted from the observed test data should be substantially larger than that from the random data. This method is recommended for factor selection (Ledesma & Valero-Mora, 2007; Hambleton & Rovinelli, 1986).

The assessment of statistical structure of test data, along with the substantive judgment, is of crucial importance in test development, quality control, and item pool maintenance. Test multidimensionality can be attributed to either dimensionally homogeneous subset of items as designed in test specification or other unexpected construct-irrelevant effects in practice (Tate, 2002). Suppose the test is designed to measure various content areas or cognitive skills. An empirical test for multidimensionality provides evidence for test validity if the internal structure of the test data conforms to the framework on which the test is built (AERA, APA, & NCME, 1999). On the condition that content or skills are moderately or highly correlated, the single IRT ability estimates can be statistically considered as a measure of a composite of the multiple abilities that the test intends to measure (Reckase, 1979; Wang, 1987, 1988). On the other hand, the inclusion of factors that are irrelevant to the target composite of abilities may introduce item bias threatening test fairness.

2.3 Item Parameter Drift

In the 1980s, researchers introduced the concept of Item Parameter Drift (IPD) to represent the changes in item parameters over time. Goldstein (1983) developed a general framework of measuring relative changes over time for repeated use of tests, while Mislevy (1982) proposed a five-step procedure to account for item parameter drift.

Earlier studies have investigated the variation of item parameters related to context effect (Eignor, 1985; Kingston & Dorans, 1984; Yen, 1980), sample statistics (Cook, Eignor, & Taft, 1988), and also the interaction of item parameter estimation procedures and sample statistics (Stocking, 1988).

Bock, Muraki, & Pfeiffenberger (1988) found one potential source of IPD was curriculum. A fourth-grade science test item about the metric system was found to be closely associated with the coverage of instruction. The time teachers spent in teaching the metric system was longer than that spent in teaching the American system. This resulted in declining difficulties for items about metric system but increasing difficulty for those about the American system. Bock et al. (1988) suggested that changes in education, technology, and culture might lead to IPD during the useful life of the scale. They found the relationship between item content and relative direction of drift, which could be attributed to a shift in Physics curricula. Similar studies were also conducted in the field of applied psychology. Chan, Drasgow, & Sawin (1999) found the effect of time on the effectiveness of the Armed Services Vocational Aptitude Battery and concluded that some cognitive-ability measures were more susceptible to time impact when compared to other item types.

During the past decade, researchers have been concerned with finding ways to identify the potential drift in item parameters. The methods include confirmatory factor analysis (Schulz, 2005), analysis of covariance models (Sykes & Ito, 1993; Sykes & Fitzpatrick, 1992), and restricted item response models (Stone & Lane, 1991). A time-dependent IRT model was used to estimate the trend over time for item parameters (DeMars, 2004; Bock, Muraki, & Pfeiffenberger, 1988).

As suggested by Angoff (1988), differential item functioning (DIF) methodology can be applied to a wide variety of important educational and psychological contexts including time, culture, geography, nationality, age, language, sex, and curricular emphasis. Similar statistical procedure can be used to assess both. Procedures for detecting item bias across subgroups were also used for detecting IPD (Donoghue & Isham, 1998; Smith, 2004). However, the simulated data only covered two time points one year apart, which was also commonly used in other studies (Wollack, Sung, & Kang, 2006). It is likely that two time points might not be sufficient to examine IPD and procedures should be generalized to multiple time points. In reality, it is typical to expect IPD over multiple testing occasions (Wollack et al., 2006).

A review of previous literature shows that IRT models have been widely used for parameter estimation or test of IPD. The validity of IRT-based techniques may deteriorate by the degree to which data meet the assumption of the model (Oshima, Raju, & Flowers, 1997). Unidimensionality is one of the principal assumptions, that is, a single construct or trait is measured by a set of items. This assumption has often been violated due to the multidimensional nature of common test purposes and items particularly for educational and psychological tests. Additionally, it is possible that the sensitivity of an item to changes in student performance on one construct area versus another may vary across years but the effect may cancel out and result in average drift close to zero. There is a lack of empirical research about the potential consequences of violation of unidimensionality upon the drift of IRT-based item parameter estimates.

Most studies simulated the variation of item parameters across time. Some research considered drift in item difficulty only (Davey & Parshall, 1995; Sykes &

Fitzpatrick, 1992; Sykes & Ito, 1993). Others identified changes in both item difficulty and item discrimination (DeMars, 2004; Wells, Subkoviak, & Serlin, 2002; Chan et al., 1999). Even though three-parameter logistical models were proposed, the lower asymptote values were fixed. Wells et al. (2002) only evaluated positive effects with the increase of the discrimination parameter by 0.5 and the difficulty parameter by 0.4. Inferences about examinees were found to be robust relative to moderate amounts of variation. Demars (2004) simulated item difficulties increased or decreased by 0.05, 0.1, 0.15, and 0.2 and the item discrimination either increased or decreased by 0.2 or 0.4. However, there was indication as to how these values were selected and whether they could be expected in the real-life datasets.

As a result, item responses were simulated to represent the null distribution of no drift condition given the ability distribution in this study. The estimates based on random samples from the real test data were considered as the alternative distribution. Significant difference between simulated data and real data suggest the existence of drift if IPD existed. In addition, a comparison of analyses using unidimensional and multidimensional IRT models was necessary for detection of IPD and impact of multidimensionality on IPD was further explored. Item parameter drift may pose a threat to the measurement of the underlying construct. Drift of anchor item parameters in particular might severely jeopardize a fair score conversation in linking items over multiple administrations, which could lead to false decisions in certification and licensure test. It has been of critical importance for the test companies involved in certification and licensure testing to guarantee that the item parameter estimations remain stable over time and across different groups of examines.

This study was conducted in two phases. The first phase explored the dimensionality and checked the measurement invariance over time. The second phase investigated item parameter drift in the context of a large-scale certification test over multiple years using both unidimensional and multidimensional techniques. A variety of test structures were compared by analyzing different combinations of sections from the real test data so as to identify the impact of multidimensionality on IPD detection.

First, data were simulated using four models with different dimensional structure. The "true" item parameters were calibrated from a "population" of examinees who took the test within these three years. The ability distributions of the samples for each year were used for simulating item responses for that year in particular. Since they were simulated from the same set of item parameters, no drift should be expected over time except for measurement error given the distribution across each sample. Compared to simulation data, the IRT-based item parameters from the real test data ought to be indifferent to the method of scaling and to the samples used in the scaling of their response data and to the year the items were administered.

CHAPTER 3

METHODOLOGY

3.1 Data

3.1.1 Instrument

The instrument used in this study is the Examination for the Certificate of Proficiency in English (ECPE), which is a large-scale certification test of English as a Foreign/ Second Language in English (EFL/ESL) designed for individuals with advanced English language ability. The test is administered annually at approximately 125 authorized testing centers in 20 countries. A new form is developed every year. No set course, syllabus, or programs of English language are required in preparation of ECPE. Candidates are encouraged to take a practice test or learn from a list of published study materials to get familiar with the format and the level of the test. Those candidates who pass all sections are awarded certificates that are recognized as official documentary evidence of advanced proficiency in the English language. The ECPE certificates can be used to apply for college or universities, to teach English, and to perform civil service activities. However, the certificates should not be used as documentation of achievement for pedagogical or training purposes.

The test consists of five separately timed sections: Interactive Oral Communication (IOC), Writing, Listening, Cloze, and Grammar/Vocabulary/Reading (GVR). The first two sections each include one task while the other three sections are multiple-choice items. Since 2003, the cloze section has been combined with the last section to create a grammar/ cloze/vocabulary/ reading (GCVR) section. All the test forms follow the same clearly specified standardized procedures for each administration. The numbers of items administered for multiple-choice sections are 40 for Listening, 40 for Cloze, and 100 for GVR, which has changed to 50 Listening items and 120 GCVR items since 2003. The IOC and Listening sections are designed to measure oracy skills while the other sections including the Writing and GCVR are designed to measure literacy skills. Each section contains subparts with different item formats. The listening section includes three types of items based on conversational exchanges, short questions, and extensive talks. The GCVR section is divided into four parts to measure understanding of grammatical structure, ability to fill in cloze in prose texts, knowledge of vocabulary common in academic and business discourse and comprehension of reading passages.

In each test form, a combination of items which have been administered earlier is inserted for the purpose of linking current test forms to previous ones. Only listening, grammar and vocabulary items are selected as linking items. Cloze and reading items are related to passages and should not be exposed as anchor items due to security. The anchor items are selected according to the item statistics in previous test forms. These items are among those having highest discrimination values and covering a wide range of difficulty levels. The items should not have any evidence of differential item functioning across samples of gender, ethnicity, and language background.

In this study, the data from forms administered in 1999, 2001, and 2004 were used. They were respectively labeled as Year 1, followed by Year 3 and Year 6, respectively. For these three forms, the same set of 30 linking items was inserted into the existing tests and scored for equating to test scores of other forms. The common items fell into three sections: 10 in listening (L), 10 in grammar (G), and 10 in vocabulary (V). There were no blank responses for these 30 common items. Correct responses were scored as "1" and wrong answers were scored as "0". The total number of examinees was 72,277, including all three administrations. Images in this dissertation are presented in color.

3.1.2 Dimensionality

Examination of the internal structure of test data can identify the dominant factors and provide evidence for hypothesized multidimensionality. Factor analytic techniques have been widely used to determine the dominant factors. However, guessing cannot be corrected in common exploratory and confirmatory factor analysis models. Exploratory analyses using a PROMAX oblique rotation of loadings were conducted in both TESTFACT (Wilson, Wood, Gibbons, Schilling, Muraki, & Bock, 2003) and NOHARM (Fraser, 1993) programs. Guessing parameters were fixed at the values estimated by BILOG-MG 3.0 (Zimowski, Muraki, Mislevy, & Bock, 2003) and then submitted to both TESTFACT (Wilson, et. al, 2003) and NOHARM (Fraser, 1993) programs for calibration.

The first step was to compute the eigenvalues based on the tetrachoric correlations using TESTFACT (Wilson, et. al, 2003) to identify important factors. In addition to the absolute values of eigenvalues, the relative changes in eigenvalues for consecutive factors were proposed by Hattie (1985) to determine the number of dominant factors. The ratio of difference in consecutive factors, denoted as Factor Difference Ratio Index (FDRI), reflected the relative change in eigenvalues. Hattie (1985) suggested the ratio of the difference between the first and the second factor and the difference between the second and the third could be used to check the relative strength of the first factor. A rule of thumb proposed by Hattie (1985) indicated that a ratio greater than three was considered as a large difference in contribution of the factors between the first and the others.

The second step consisted of assessing the dimensionality by fitting multidimensional models of varying solutions and assessing the fit by residual statistics. Different statistics were computed by both TESTFACT (Wilson, et. al, 2003) and NOHARM (Fraser, 1993). They were similar in reflecting the difference between the observed and model-based relationship between items. The Root Mean Square Error of Approximation (RMSEA) values using TESTFACT (Wilson, et. al, 2003) were no greater than 0.05 indicating a close approximation between observed and expected values. Alternatively, the Root Mean Square of Residuals (RMSR) was based on the difference between the observed item correlations and those implied by models and should be 0.05 or less for an acceptable factor solution (Muthen & Muthen, 2001). The RMSRs were also available in output for NOHARM (Fraser, 1993) but the residual statistic was based on the difference between the observed and model-based proportions of correct response for each pair of items. As suggested by McDonald and Mok (1995), the criteria was equal to or less than four times the reciprocal of the square root of the sample size indicating fit to the data. The Tanaka (1993) index greater than the criteria value of 0.95 also indicated a good model fit to the data.

For detecting full dimensionality, DIMTEST procedures (Stout, 1987; Stout, Froelich, & Gao, 2001) were used to test whether the internal structure of a selected set of items (called AT items) was homogeneous but distinctive from other items (called PT items). The theoretical hypothesis was that the covariances among AT items conditional on PT item scores were zero when fitting a undimensional model but positive for items measuring more than one dimension.

Parallel analyses (Ledesma & Valero-Mora, 2007; Horn, 1965) were subsequently applied to determine the number of factors to retain. Eigenvalues based on random variables were generated with zero correlation were compared to those computed from the observed item responses. The simulated dichotomous data were selected from the binomial distributions with the means equal to the observed proportion of correct response for each item from real data. Even though both the real data and simulated data had the same number of items coded as one, a distinct factor could be extracted if the eigenvalues were larger than those obtained from the random uncorrelated data ((Ledesma & Valero-Mora, 2007). The number of dimensions was then determined by the factors with eigenvalues greater than that expected from the 400 replications of random uncorrelated data and precedes a significant drop in a scree plot.

Finally, plots of the item vectors were used to decompose the 30 linking items into essential dimensionalities. Proposed by Reckase and Ackerman (Ackerman, 1994; Reckase, 1985), item vector plots were scatterplots of item difficulties using an oblique factor analysis solution with three factors. The lengths of vectors were proportional to the multidimensional discrimination while the origin of the vectors indicated the three-dimensional item difficulties. When item vectors fell on a straight line, one essential domain was identified.

Previous research showed that the whole test had a dominant factor that is overall English skill. However, factor analyses also showed a clear pattern of structure with items within the same section tending to load high on the same factor except for a few cases (Johnson, Li, Yamashiro, & Yu, 2006a; 2006b). In this study, the data were from responses to the linking items representing English proficiency in grammar, listening, and vocabulary. By analyzing arbitrary combinations of sections in terms of the linking items, a variety of dimensionality structures were tested and the effects of these different dimensionality structures on IPD were explored.

3.1.3 Measurement Invariance

The establishment of measurement invariance has become a logical prerequisite before conducting the substantive analysis across groups (Vandenberg & Lance, 2000). For the purpose of checking different dimensional structures, the measurement models included the one factor model, two factor model, and three factor model. A confirmatory factor analysis (CFA) was used to test the factor structure underlying a set of measurements (Kaplan, 2000). Unmeasured covariances were assumed for each pair of latent variables but there were no direct paths connecting the latent variables.

The primary analytic tool was structural equation modeling using Mplus software (Muthen, & Muthen, 2001). Θ parameterization was also used so that residual variances for continuous latent response variables were allowed to be parameters in the model. Because the data were dichotomous for each item, this was appropriate for models where a categorical variable was influenced by or influenced another observed or latent dependent variable (Muthen & Muthen, 2001).

Model fit was determined by several indices. The Comparative Fit Index (CFI) was used (Bentler, 1990). CFI values close to one indicated a very good fit. The Tucker-Lewis index (TLI), also called the Bentler-Bonett non-normed fit index (NNFI), was not necessarily vary from 0 to 1. The values close to one suggested a good model fit to the data. Another index used was the Root Mean Square Error of Approximation (RMSEA), where a value of zero indicated perfect model fit (Browne & Cudeck, 1993). Chi-square difference was used to test whether the improvement of model fit is statistically significant.

3.2 Design

The item parameter estimates from concurrent calibration of the entire population including all three years of administration were taken as "true" item parameter values. The mean and variance-covariance matrix for abilities were then estimated with the item parameter fixed. Based on these "true" item parameters and ability distribution of each year, simulated data were generated using the three-parameter three-dimensional logistic model. A sample of responses for 2,000 examinees was simulated for each administration year and this process was replicated 400 times. The distributions of these simulated data involved items with the same parameters and represented the null hypothesis with no drift. In addition, a random sample of 2000 examinees from the real data was selected without replacement for each administration year, which represented the distribution of the alternative hypothesis for testing the IPD.

In order to reveal the effects of multidimensionality on item parameter drift, four combinations of dimensional structures were examined, as outlined in Table 1, for the same set of common items. The assumed dimensionality for each calibration reflected a certain factor structure. In the beginning, all 30 items in three sections were combined together and taken as one-dimensional model measuring English language ability. Secondly, sections of grammar and vocabulary were referred to as literacy skills in language assessment while listening is considered as oracy skills. They were scored as two subscales but each scale was essentially one-dimensional. For the third condition, the three sections were considered as independent from each other and calibrated separately. Each section was truly unidimensional as designed and was measuring a particular area of English ability. Based on the three-parameter logistic IRT model, the first three calibrations were run using PARSCALE 4.1 (Muraki & Bock, 2003). For comparison, an underlying three-dimensional solution was assumed for model IV. Item parameters were estimated using TESTFACT 4.0 (Wilson, et. al, 2003). As a result, there were 24 sets (4 calibrations * 3 years * 2 data) of 400 item parameters estimates after replication.

Table 3. 1 Combinations of item calibration

Model	Type	Subscale	Dimensionality	Model	Item parameter per item	Software
I	LGV	NA	One	3 PL	c, a, b	PARSCALE
II	L,GV	Two	One	3 PL	c, a, b	PARSCALE
III	L,G,V	Three	One	3 PL	c, a, b	PARSCALE
IV	LGV	NA	Three	3-D 3PL	c, a1, a2, a3, d	TESTFACT

Note: L refers to listening items; G refers to grammar items; V refers to vocabulary items; 3 PL refers to three-parameter logistic model; 3-D 3 PL refers to three-dimensional three-parameter compensatory logistic model. NA means not applicable.

Because of the indeterminacy and the way different programs set the origin, units, and orientation of the axes, item parameter estimates from different calibrations needed to be placed on the same scale using a linking process. For the unidimensional model, the item parameters have a linear relationship. As a result, a linear transformation was necessary to place the item parameter estimates on a common scale. The scaling used the characteristic-curve method (Stocking & Lord, 1983). Under this approach, estimated "true scores" were equated using least squares. The base scale was set by the concurrent calibration of all three administration years as large samples result in a smaller sampling error other things being equal (Oshima et al., 1997). Calibrations of all replications underlying the first three models were converted to these base scales.

Indeterminacies are also important when calibrated item parameters under multidimensional theory. Three types of indeterminacy were summarized by Li and Lissitz (2000). In the coordinate system, both the point of origin and the unit along the axes are undefined. The MIRT parameter estimation program TESTFACT (Wilson, et. al, 2003) addresses these identification problems by setting the estimated proficiencies to be distributed as a multivariate normal with a mean vector of zero and the identity matrix for the variance-covariance matrix. The unit was the standard deviation of the observed proficiencies (Li et al., 2000). The third type of indeterminacy was due to the orientation of the coordinate system. TESTFACT (Wilson, et. al, 2003) addresses this issue by setting the coordinate system to be orthogonal, which sets the correlations among the coordinates to zero. Li and Lissitz (2000) proposed an approach to computing a composite transformation for changing the linked group's reference system into the base group's reference system by an orthogonal Procrustes rotation, a translation transformation, and a single dilation. An extension of these methods to a more general approach used the oblique Procrustes method (Mulaik, 1972) based on the work by Martineau (2004) and Reckase and Martineau (2004). In Reckase (2006), the rotation matrix was defined as:

$$\mathbf{Rot} = (\vec{\mathbf{a}}_{\mathbf{a}}{}'\vec{\mathbf{a}}_{\mathbf{a}})^{-1}\vec{\mathbf{a}}_{\mathbf{a}}'\mathbf{a}_{\mathbf{b}}, \tag{3.1}$$

where $\vec{a}_a$ is a n×m matrix of discrimination parameters for the reference system of the linked group and $\vec{\mathbf{a}}_{\mathbf{b}}$ is a n×m matrix of discrimination parameters for the reference system of the base group. **Rot** is the m ×m rotation matrix for the discrimination parameters.

The rotated **a**-matrix to the base scale for the linked group is thus given by:

$$\hat{\mathbf{a}}_{\mathbf{b}} = \vec{\mathbf{a}}_{\mathbf{a}}\mathbf{Rot} \tag{3.2}$$

Accordingly, the *d*-parameters can be rescaled by adding the transformation matrix as follows:

$$\hat{\mathbf{d}}_{\mathbf{b}} = \vec{\mathbf{d}}_{\mathbf{a}} + \mathbf{TRAN} = \vec{\mathbf{d}}_{\mathbf{a}} + \mathbf{a}_{\mathbf{a}}(\vec{\mathbf{a}}_{\mathbf{a}}{}'\vec{\mathbf{a}}_{\mathbf{a}})^{-1}\vec{\mathbf{a}}_{\mathbf{a}}'(\vec{\mathbf{d}}_{\mathbf{b}} - \vec{\mathbf{d}}_{\mathbf{a}}), \tag{3.3}$$

where $\vec{\mathbf{d}}_{\mathbf{b}}$ is a $n\times 1$ vector of d-parameters for the reference system of linked group and $\vec{\mathbf{d}}_{\mathbf{a}}$ is a $n\times 1$ vector of d-parameters for the reference system of base group. **TRAN** becomes the $m \times 1$ transformation vector for the *d*-parameters.

3.3 Analysis

Item parameter estimates converted to the same scale across different administration years were compared first for replications of simulated samples and then samples from real data. The means of parameter estimates of the 400 replications for the common items were plotted to detect significant discrepancy over time. Such plots can show the deviation in distributions of the item parameter estimates for real data samples compared to those from simulation samples. The simulation data were assumed to have no drift in parameter estimates because they were generated using the same set of item parameters. Items that showed the most aberrant deviation over time in the real data samples might reveal drift. Invariant item parameter estimates also suggested good model/test response data fit.

Even if differences were observed for parameter estimates of the common items, it was necessary to test whether these differences were statistically significant or are simply due to random error. The standard detection method for differential item functioning (DIF) could also be applied to detection of IPD. An extension of the method for differential item and test functioning (DFIT), developed by Raju, van der Linden, and Fleer (1995), was used here to study IPD. This framework compared test characteristic curves and could be applied to either unidimensional or multidimensional tests (Oshima, Raju, & Flowers, 1997), as follows.

The probability of correctly answering item *i* for examinee *j* based on the three-parameter logistic IRT model (Lord, 1980) is given by Equation 2.1. The probability of correctly answering item *i* for examinee *j* based on the multidimensional three-parameter compensatory logistic model (Reckase, 1985; Reckase & McKinley, 1991) is given by:

$$P_i(Y_{ij} = 1 \mid \vec{\mathbf{a}}_{\mathbf{i}}, c_i, d_i, \vec{\boldsymbol{\theta}}_{\mathbf{j}}) = c_i + (1 - c_i)\frac{\exp(1.702\vec{\mathbf{a}}_{\mathbf{i}}'\vec{\boldsymbol{\theta}}_{\mathbf{j}} + d_i)}{1 + \exp(1.702\vec{\mathbf{a}}_{\mathbf{i}}'\vec{\boldsymbol{\theta}}_{\mathbf{j}} + d_i)}, \tag{3.4}$$

where $\vec{\mathbf{a}}_{\mathbf{i}}$ is a $m \times 1$ vector of item discrimination parameter estimates for item *i*, d_i is a scalar parameter representing item difficulty for item *i*, c_i is the lower asymptote parameter for item *i*, $\vec{\boldsymbol{\theta}}_{\mathbf{j}}$ is a $m \times 1$ vector of the ability parameters for examinee *j*, and m refers to the number of dimensions for ability parameters. The scaling constant of 1.702 accounts for the difference between the logistic function and normal ogive function.

IRT-based true scores are given as:

$$\tau_j = \sum_{i=1}^{k} P(Y_{ij} = 1 \mid a_i, b_i, c_i, \theta_j) \ , \tag{3.5}$$

for the unidimensional model and

$$\tau_j = \sum_{i=1}^{k} P(Y_{ij} = 1 \mid \vec{\mathbf{a}}_{\mathbf{i}}, d_i, c_i, \vec{\boldsymbol{\theta}}_{\mathbf{j}}), \qquad (3.6)$$

for the multidimensional model.

Assuming the examinees' true score is independent of group membership, the differential test functioning (DTF) is defined by Raju, van der Linden, & Fleer (1995) as:

$$DTF = E_F(\tau_F - \tau_R)^2 = E_F D^2 = \sigma_D^2 + (\mu_{\tau F} - \mu_{\tau_R})^2 = \sigma_D^2 + \mu_D^2 \qquad (3.7)$$

where E is the expectation taken over either the reference group or the focal group, μ and σ refer to the mean and standard deviation for each group, and D is given by $\tau_F - \tau_R$.

The equation shown above suggests the compensating nature of the proposed DTF. The difference in probability of one item for the focal group compared to the reference group is canceled out by the difference in another item probability at the test level.

To represent the potential compensating drift at the item level, nonconfirmatory differential item functioning (NDIF) assumes that all items in the test are free of DIF except for the item examined, which corresponds to most of the IRT-based DIF methods. NDIF is expressed as (Raju, et al., 1995):

$$NCDIF_i = E_F(P_{iF}(\theta) - P_{iR}(\theta))^2 = E_F d_i^2 = \sigma_{d_i}^2 + \mu_{d_i}^2, \text{ for } d_j = 0, \text{ and } j \neq i, \qquad (3.8)$$

where P_{iF} and P_{iR} are the probability of a correct response at a given θ value (or vector for multidimensional model) using item parameter from the reference group and the focal group, respectively, and d_i refers to the difference in probability for item i for the same examinee. The relationship between D and d is: $D = \sum_{i=1}^{k} d_i$, and D is the true score difference for an examinee. However, only estimates are available in practice to compute these indexes. The NCDIF is estimated for each calibration set of each item. The distribution of 400 replications for the simulation data serves as the null distribution while the distribution of the one based on real data is for the alternative hypothesis.

In this study, the null distribution of NCDIF is generated by sorting the 400 NCDIF values calculated using the simulation data. As assumed, the simulation data represent the no drift situation except for measurement error. A cut-off value is then determined by obtaining a 100(1- α) percentile with the type I error rate of α. Given the choice of α values of 0.05 or 0.01, the 95% and 99% confidence intervals are computed for the distribution of the NCDIF index.

Out of the 400 replications, the count of NCDIF index values that were greater than the cut-off values suggests deviation of the distribution of NCDIF for real data from the distribution based on simulation data. The null hypothesis was that the NCDIF distributions from real data samples were equivalent to that from the. simulation samples. The larger the number of index values out of 400 replication, the more frequent the NCDIF values in the real data sample were rejected as being the same distribution as the null.

CHAPTER 4

RESULTS & DISCUSSIONS

This chapter has three parts. The first part summarizes the descriptive statistics and dimensionality analysis results for the linking items of a large-scale certification test in English as a Foreign/Second Language Assessment. The second part reviews the invariance property of the item parameters. The last part compares the results for item parameter drift and statistical tests for significance for four models.

4.1 Descriptive Statistics

The descriptive statistics shown in Table 4.1 include means and standard deviations for items and scales scores for items in each section, based on the number of correct responses. It was observed that the number-correct score means of these items for examinees at Year 6 were consistently lower than those of the previous two years for the total scores and scores of each section. Kuder-Richardson Formula 20 (KR20) estimates of reliability, known as a special case of Cronbach's alpha, were used for dichotomies in particular. That is, the items were scored as “1” for correct responses and “0” for wrong responses. Similar estimates of reliability were found to be above 0.7 across years of administrations, which was lower than the criterion value of 0.9 for a homogenous test. The KR-20 is known to be a function of item difficulty, spread in test scores and test length. The values in this study, however, might be underestimated because only a small part of the test (linking items) was examined in this study. Also reported are the number of examinees who were administered the test each year.

Table 4. 1 Descriptive statistics and scale reliability

Year	No. of Items	Case	Total			Listening		Grammar		Vocabulary	
			KR 20	Mean	*S.D.*	Mean	*S.D.*	Mean	*S.D.*	Mean	*S.D.*
Year 1	30	17151	0.74	20.38	4.54	7.65	1.88	7.09	2.06	5.65	2.11
Year 3	30	22099	0.72	20.81	4.31	7.73	1.78	7.08	2.02	6.01	2.01
Year 6	30	33027	0.74	19.86	4.62	7.58	1.89	6.75	2.07	5.54	2.15
Total	30	72277	0.74	20.28	4.53	7.64	1.85	6.93	2.06	5.71	2.11

Note: S.D. refers to standard deviation; KR 20 refers to Kuder-Richardson Formula 20.

4.2 Dimensionality

Examination of the internal structure of test data can be used to identify the dominant factors and to provide evidence for hypothesized multidimensionality. Exploratory analyses using a PROMAX oblique rotation of loadings were conducted in both TESTFACT and NOHARM programs because they could model lower asymptote in parameter estimation. Guessing parameters were fixed at the values estimated by BILOG-MG and then submitted to both TESTFACT and NOHARM programs for calibration. A series of factor solutions were estimated on the basis of one-, two-, and three-factor models by both programs. The first step was to compute the eigenvalues based on the tetrachoric correlations to identify important factors. The second step consisted of assessing the dimensionality by fitting multidimensional models of varying solutions and assessing the fit by residual statistics. Finally, scree simulation plots and plots of the item vectors were compared along with model fit statistics to decompose the 30 linking items in terms of the essential dimensionality.

The leading eigenvalues from the tetrachoric correlation matrix computed by TESTFACT are shown in Table 4.2. There was no output of eigenvalues for NOHARM because the program analyzed the sample proportion correct for item pairs instead of a tetrachoric correlation matrix. The first three factors all had eigenvalues that were greater than one. The first factor was dominant with eigenvalues that were around seven, and the second factor was strong with eigenvalues that were close to three.

The ratios of factor differences were also computed and presented in Table 4.2. The datasets met the criterion value of three except for year 3 but the value was very close to three, which confirmed that the first factor dominated while the other factors from the second onward made relatively minor contributions.

As would be expected, the eigenvalue analysis verified the existence of a dominant factor and two minor factors for the test data. The strong first factor extracted accounted for approximately 21% of the total variance for year 1, 20% for year 3, and 22% for year 6. However, the second factor accounted for no more than 8 % of variation and the third factor accounted for less than 4% of the variance. Reckase (1979) suggested the first factor accounting for at least 20 % of total variance verified a dominant underlying latent factor for items concerned. The percentages associated with the first factor were close to critical values implied an approximation to unidimensionality for the test data.

Table 4. 2 Eigenvalues, FDRI, and percentage of variance explained using TESTFACT

Year	Eigen Values			FDRI	Percentage of Variance Explained		
	F 1	F 2	F3		F 1	F 2	F3
Year 1	7.2677	3.0304	1.8943	3.7296	20.88%	7.51%	3.94%
Year 3	6.7886	2.9947	1.6245	2.7689	19.58%	7.40%	3.26%
Year 6	7.7298	2.8272	1.6261	4.0819	22.50%	6.87%	3.26%
Total	7.3029	2.8712	1.6578	3.6523	21.14%	6.99%	3.32%

Note: FDRI refers to Factor Difference Ratio Index.

Residuals and fit statistics were also compared for one through three factor models and were summarized in Table 4.3. Different statistics were computed by each program but they were similar in reflecting the difference between the observed and model-based relationship between items. Compared to the criteria of 0.05, the Root Mean Square Error of Approximation (RMSEA) values using TESTFACT were no greater than 0.03 for each factor solution indicating a close approximation between observed and expected values. Alternatively, the Root Mean Square of Residuals (RMSR) was based on the difference between the observed item correlations and those implied by models and should be 0.05 or less for an acceptable factor solution (Muthen & Muthen, 2001). A hypothesized one-factor model, resulting in RMSRs that were close to 0.09, did not provide support for a good fit to the data. However, the values decreased substantially to approximately 0.05 for two-factor model and around 0.03 for three-factor model. Comparison of these results with those for the one-factor model suggested that three-factor solution provided the best fit to the data.

The RMSRs were also available in output from NOHARM but the residual statistic was based on the difference between the observed and model-based proportions of correct response for each pair of items. As suggested by McDonald (1991), the criteria was equal to or less than four times the reciprocal of the square root of the sample size indicating fit to the data. The criteria values computed for this study was 0.031 for year 1, 0.027 for year 3, 0.022 for year 6, and 0.015 for three years combined. A gradual decrease in RMSRs was observed across the three solutions. The Tanka indexes were also in generally greater than the criteria value of 0.95 indicating good model fit. Index values for the three-factor model solution were close to one, which suggested a nearly perfect fit to the data. The appreciable improvement of model fit occurred after adding the second and the third factors.

Table 4. 3 Goodness-of-fit statistics for TESTFACT and NOHARM exploratory solutions

Statistics from TESTFACT by year	Root Mean Square Error of Approximation (RMSEA)			Root Mean Square of Residuals (RMSR) [a]		
	F 1	F 2	F3	F 1	F 2	F3
Year 1	0.0270	0.0267	0.0266	0.0942	0.0554	0.0330
Year 3	0.0227	0.0224	0.0224	0.0871	0.0461	0.0332
Year 6	0.0190	0.0187	0.0187	0.0829	0.0476	0.0339
Total	0.0117	0.0115	0.0115	0.0902	0.0519	0.0317
Statistics from NOHARM by year	Root Mean Square of Residuals (RMSR) [b]			Tanaka index of goodness of fit		
	F 1	F 2	F3	F 1	F 2	F3
Year 1	0.0073	0.0043	0.0026	0.9799	0.9930	0.9974
Year 3	0.0065	0.0034	0.0023	0.9760	0.9934	0.9969
Year 6	0.0067	0.0036	0.0025	0.9797	0.9941	0.9973
Total	0.0066	0.0036	0.0023	0.9801	0.9941	0.9975

Note: [a] RMSR in TESTFACT were based on the residual correlations as the difference between model-based and observed item correlations.

[b] RMSR in NOHARM were based on the residual correlations as the difference between model based and observed proportions of correct responses.

Since the residuals with large absolute values could be canceled out by taking the average, the number of residual with absolute values that were greater than 0.01 were counted and shown in Table 4.4. For the first year, the residual correlation matrix obtained from TESTFACT output displayed a substantial decrease from 107 cases to 3 cases as the number of factor increased for Year 1, from 49 to 1 for Year 3, and from 50 to 1 for Year 6. The residuals with large absolute values also decreased from 49 to 1 when the data for all three years were analyzed together. The results from the output in NOHARM also demonstrated a similar pattern. Dramatic decreases for the residuals with large absolute values were observed for the three-factor solution suggesting extensive improvement of the model fit to the data.

Table 4. 4 Number of residual with absolute values greater than one comparing both TESTFACT and NOHARM exploratory solutions

Year	TESTFACT			NOHARM		
	F 1	F 2	F3	F 1	F 2	F3
Year 1	107	13	3	64	13	3
Year 3	49	7	1	49	7	2
Year 6	50	10	1	50	10	1
Total	49	9	1	52	9	1

Even though the exploratory analysis evidenced a dominant first factor that could be attributable to most of the variability in observed scores, it was necessary to check whether the other two minor factors were essentially different from the first factor and were significant constructs. Other than the eigenvalue and residual analyses shown above, graphic tools for measures of fit were also included to compare models of different orders and to provide substantive support for multidimensionality.

The DIMTEST analyses were conducted first to test the hypothesis whether the selected items had distinct structure compared to the rest of the items. As expected, DIMEST rejected the null hypothesis of unidimensionality. The results also showed that the set of items selected by the program as AT items were all from the listening sections suggesting that at least listening items were dimensionally homogeneous but distinct from the items in the other two sections.

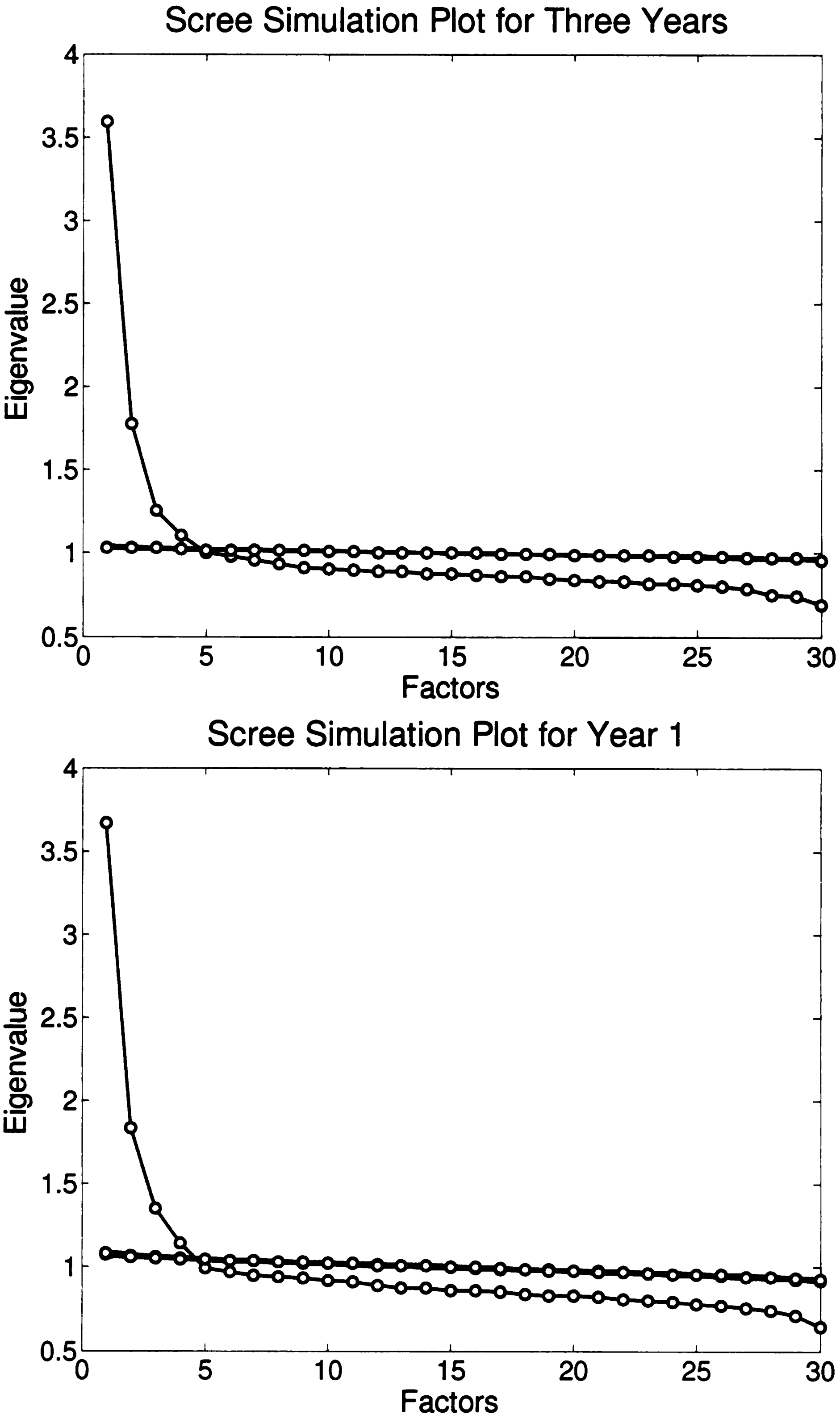


Figure 4. 1 Scree simulation plots for three years and Year 1

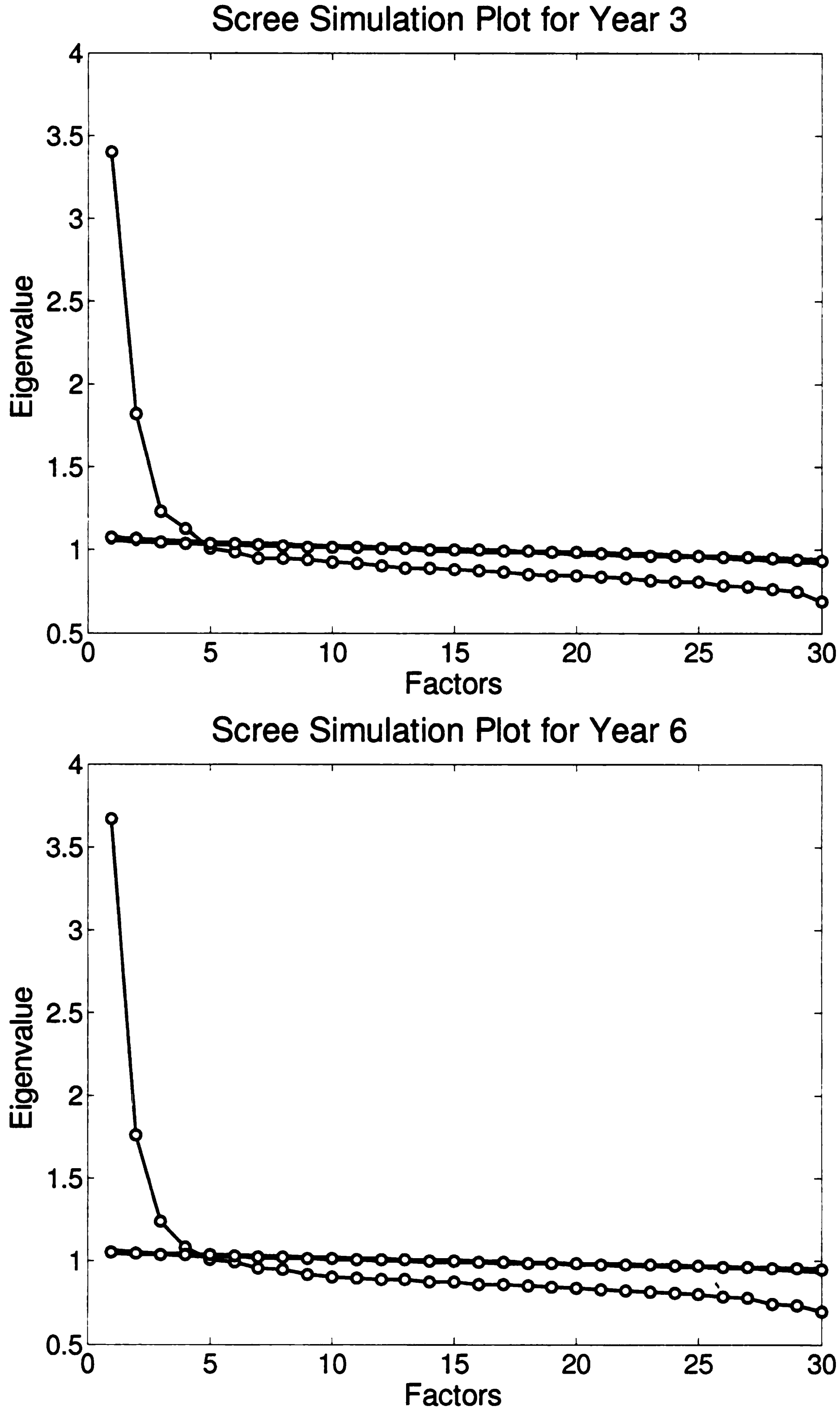
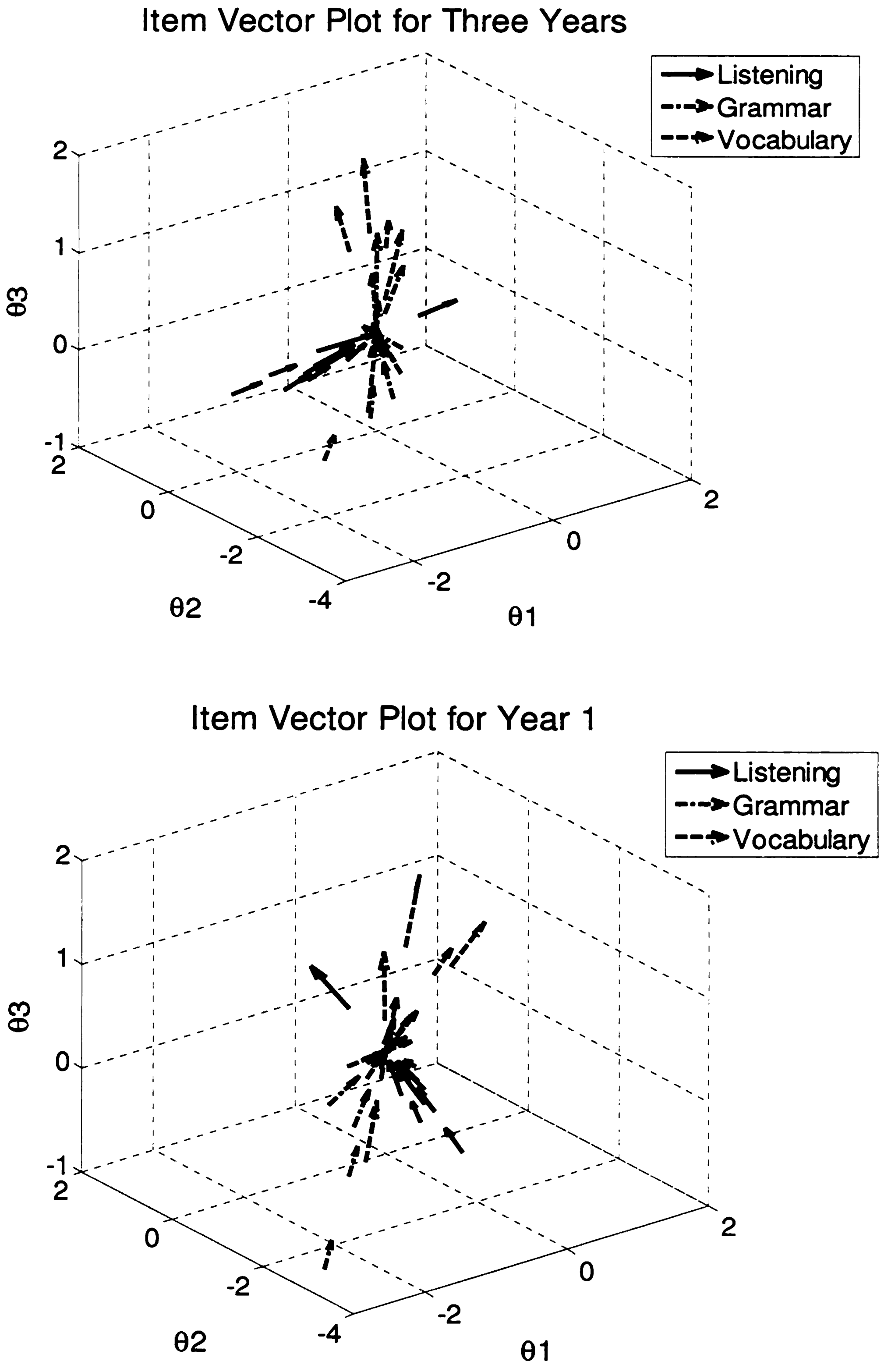


Figure 4. 2 Scree simulation plots for Year 3 and Year 6

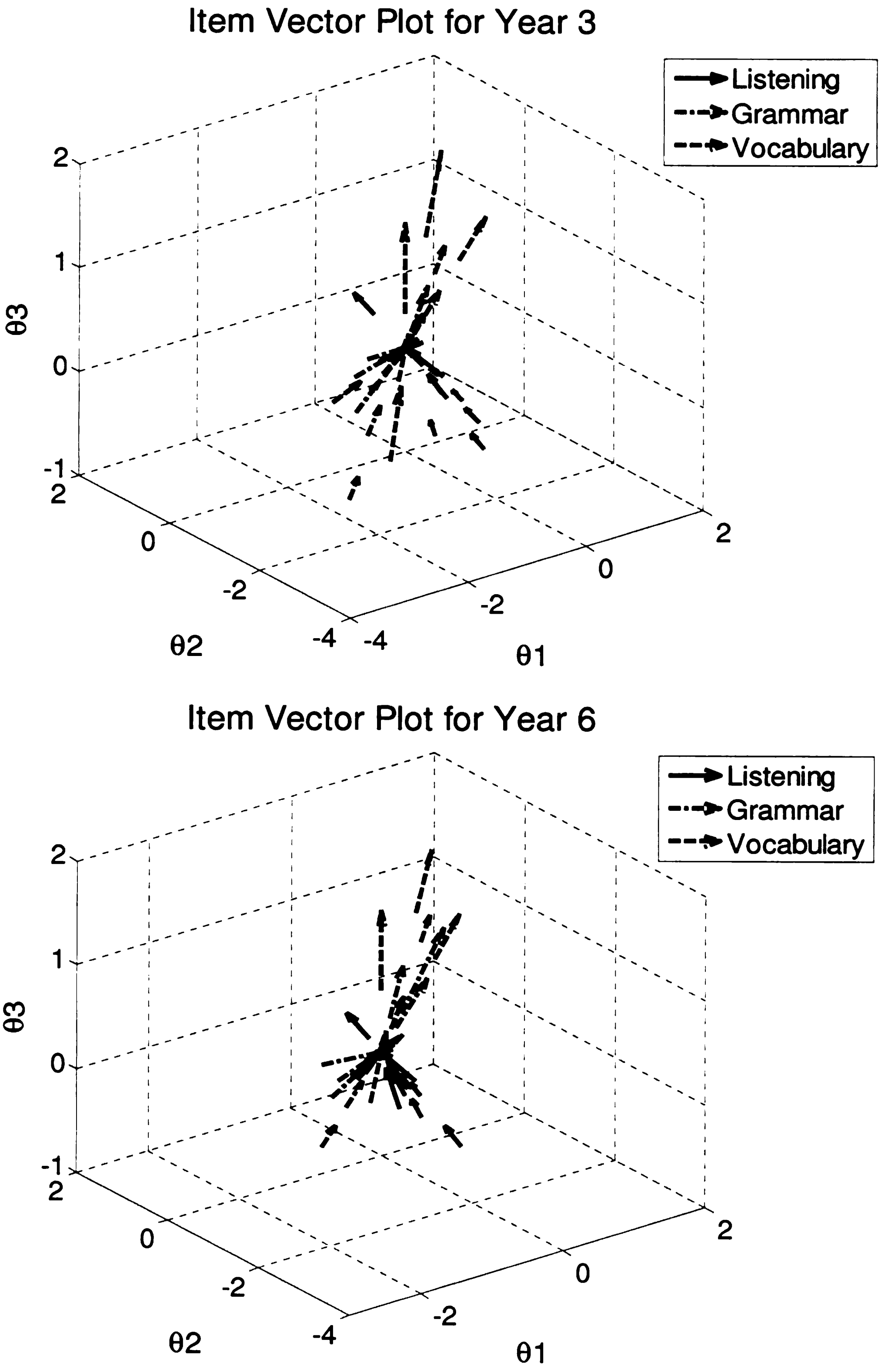
The visual outputs from the parallel analysis are presented in Figures 4.1 and 4.2. The scree simulation plots are displayed for the whole population with three years combined and the subpopulation for each year. The plots included the scree plot of the observed value based on the real data and all the scree plots resulting from the simulated data. The scree plots from the simulated data overlapped and remained a line with eigenvalue close to one. Even though four factors have eigenvalues exceeded those obtained from the randomly simulated data with zero correlation, the fourth factor was very close to the eigenvalue of one. Also, extraction of four factors did not provide a plausible interpretation for the clusters of items. However, a clear pattern of three-dimensional clustering was consistent for the three domains measured.

Figures 4.2 and 4.3 displayed the item vector plots for test data for each year and the test data with three years combined, which also provided evidence for a three-factor model across administrations. These item vectors were plotted with regard to orthogonal solutions with the three-factor model. The vector plots revealed the separation of items into more than one group. The listening items for all four plots showed a clear pattern of pointing in the same direction, which suggested they constitute an essentially unidimensional scale. However, the grammar items and vocabulary items varied widely in their orientations. In Panel A both grammar and vocabulary items mixed with each other but appear to comprise two clusters. For Panels B and D, both types of items mostly clustered at a common vector. In Panel C, most of the items were oriented in two different directions but with a small angle between them. The listening items measured the same combination of skills and the other 20 items measured two different composites of abilities.


Item Vector Plot for Three Years

Item Vector Plot for Year 1

Figure 4. 3 Item vector plots for three years and Year 1



Item Vector Plot for Year 3

Item Vector Plot for Year 6

Figure 4. 4 Item vector plots for Year 3 and Year 6

It is important that the factor solutions are interpretable for the purpose of determining the number of dimensions (Gorsuch, 1983). Each of the items had substantial loadings on one factor. The highest loadings are highlighted in bold in the tables. The highlighted items could be used as indicators to represent the factors for a three-factor solution. Inspection of the Promax rotated factor loadings, given in Table 4.5, showed that the three factors extracted were mathematically acceptable and nontrivial. Only listening items loaded high on the second factor for each year suggesting that the factor represents listening capabilities. Most of the grammar items loaded on the first factor except for a few items. The vocabulary items were separated into two groups, six of which loaded on the first factor but the other four loaded on the third factor.

The inter-factor correlations for the Promax rotated solution were given in Table 4.6. The highest correlations were greater than 0.60 between the first factor and the third factor that represented grammar and vocabulary items. This was consistent with the test design that both types of items measure English literacy skills. The first and the third factors were also moderately correlated with the second factor indicating oracy skills represented by all listening items.

In summary, a similar number of dimensions were identified by eigenvalue analyses, residual and fit statistics and graphic assessment. Although residual statistics suggested a parsimonious one-factor solution could fit the model well, graphic analyses and factor loadings implied a model with two or three factors should be considered for better identification and interpretation. The second factor was clearly represented by listening items; three fourths of the grammar and vocabulary items had the highest loadings on the first factor and one-fourth on the third factor. The third factor, much weaker than the other two, showed some evidence of representing uniqueness of vocabulary items. These tests had in essence three-dimensional components that were in agreement with the theoretical structure in terms of English proficiency in three skill areas.

Table 4.5 Promax rotated factor loadings based on three-factor model

	Total			Year 1			Year 3			Year 6		
Item	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
L01	0.05	**0.50**	0.01	0.07	**0.50**	0.05	0.03	**0.47**	0.03	0.08	**0.53**	0.02
L02	0.00	**0.49**	0.04	0.07	**0.51**	0.04	0.07	**0.50**	0.02	0.06	**0.50**	0.06
L03	0.09	**0.64**	0.00	0.20	**0.72**	0.12	0.11	**0.63**	0.05	0.07	**0.65**	0.00
L04	0.09	**0.46**	0.11	0.13	**0.45**	0.13	0.14	**0.41**	0.15	0.04	**0.48**	0.08
L05	0.10	**0.60**	0.02	0.14	**0.61**	0.03	0.10	**0.58**	0.08	0.07	**0.61**	0.06
L06	0.02	**0.39**	0.03	0.00	**0.38**	0.04	0.04	**0.30**	0.07	0.00	**0.45**	0.02
L07	0.09	**0.50**	0.13	0.22	**0.55**	0.28	0.08	**0.48**	0.10	0.04	**0.47**	0.09
L08	0.13	**0.79**	0.01	0.02	**0.71**	0.09	0.10	**0.83**	0.04	0.17	**0.77**	0.02
L09	0.10	**0.62**	0.08	0.09	**0.65**	0.08	0.06	**0.63**	0.04	0.10	**0.60**	0.07
L10	0.20	**0.42**	0.15	0.14	**0.48**	0.05	0.19	**0.46**	0.12	0.19	**0.35**	0.18
G11	**0.46**	0.06	0.12	**0.48**	0.02	0.17	**0.46**	0.05	0.12	**0.44**	0.10	0.10
G12	**0.41**	0.02	0.01	**0.28**	0.02	0.12	**0.46**	0.07	0.02	**0.43**	0.01	0.02
G13	**0.59**	0.13	0.00	**0.51**	0.11	0.12	**0.64**	0.18	0.03	**0.56**	0.10	0.02
G14	**0.41**	0.04	0.11	0.24	0.05	**0.29**	**0.40**	0.07	0.12	**0.46**	0.06	0.03
G15	**0.34**	0.23	0.04	**0.21**	0.19	0.17	**0.33**	0.21	0.03	**0.37**	0.27	0.12
G16	**0.63**	0.28	0.24	**0.83**	0.24	0.38	**0.72**	0.23	0.31	**0.51**	0.32	0.17
G17	0.29	0.03	**0.41**	0.07	0.02	**0.52**	0.28	0.00	**0.39**	**0.45**	0.04	0.30
G18	**0.86**	0.19	0.24	**0.85**	0.17	0.17	**0.80**	0.24	0.15	**0.86**	0.17	0.26
G19	**0.37**	0.32	0.04	0.24	**0.29**	0.21	**0.47**	0.22	0.04	0.35	**0.40**	0.01
G20	**0.26**	0.09	0.11	0.11	0.06	**0.29**	**0.29**	0.08	0.11	**0.31**	0.12	0.01
V21	0.16	0.19	**0.72**	0.25	0.14	**0.71**	0.09	0.23	**0.65**	0.13	0.21	**0.74**
V22	0.10	0.10	**0.57**	0.01	0.07	**0.57**	0.07	0.11	**0.55**	0.18	0.11	**0.58**
V23	0.10	0.20	**0.82**	0.20	0.18	**0.82**	0.16	0.20	**0.83**	0.01	0.17	**0.74**
V24	0.18	0.08	**0.35**	0.03	0.08	**0.48**	0.18	0.10	**0.33**	**0.27**	0.08	0.26
V25	**0.38**	0.16	0.34	**0.41**	0.15	0.29	**0.35**	0.16	0.31	0.37	0.17	**0.41**
V26	**0.46**	0.07	0.09	**0.39**	0.08	0.13	**0.43**	0.07	0.09	**0.49**	0.05	0.09
V27	**0.45**	0.02	0.06	**0.57**	0.01	0.10	**0.48**	0.02	0.08	**0.34**	0.02	0.03
V28	**0.57**	0.01	0.03	**0.51**	0.03	0.11	**0.51**	0.02	0.07	**0.58**	0.03	0.04
V29	**0.63**	0.11	0.03	**0.66**	0.10	0.03	**0.57**	0.09	0.01	**0.65**	0.16	0.01
V30	**0.29**	0.13	0.12	**0.36**	0.10	0.07	**0.30**	0.13	0.08	**0.25**	0.15	0.16

Table 4.6 Promax factor correlations based on three-factor model

	Total			Year 1			Year 3			Year 6		
Factor	Fl	F2	F3	Fl	F2	F3	Fl	F2	F3	Fl	F2	F3
F 1	1.00			1.00			1.00			1.00		
F 2	0.48	1.00		0.43	1.00		0.45	1.00		0.50	1.00	
F 3	0.64	0.41	1.00	0.66	0.39	1.00	0.64	0.38	1.00	0.60	0.39	1.00

A final check was to ensure that the model solution was generally supported across years while there was also evidence of better model fit with increasing number of factors as shown in Table 4.7. The measurement models were checked first including all examinees. The three-factor solution had the best fit indices (CFI and TLI <.95; RMSEA <.05), which remained when the models were compared by examinees of each year as a separate group. An invariant factor pattern was specified accordingly. For the purpose of testing the measurement equivalence, a weak assumption is that all factor loadings are constrained to be equal across groups and a strong assumption is both factor loadings and thresholds are equal across groups. The restriction of equal factor loadings demonstrated an improvement of model fit suggesting that loadings of like items within the invariant factor pattern were also equal over time. The additional constraint of equal threshold that separates between the correct and incorrect response showed a decrease in fit. In general, the three-factor model proved to have better fit compared to the other two solutions.

Table 4.7 Measurement equivalence compared by models and years

Model	df	Chi-square	CFI	TLI	RMSEA
Measurement model					
One-factor model	369	39119.093	0.806	0.848	0.038
Two-factor model	365	18996.715	0.907	0.926	0.027
Three-factor model	364	18305.540	0.910	0.929	0.026
Comparison over three years					
One-factor model	1082	40581.958	0.800	0.843	0.039
Two-factor model	1069	20113.427	0.904	0.923	0.027
Three-factor model	1066	19345.394	0.908	0.926	0.027
Invariance of factor loadings					
One-factor model	970	33011.232	0.838	0.858	0.037
Two-factor model	986	17841.260	0.915	0.927	0.027
Three-factor model	985	17232.063	0.918	0.929	0.026
Invariance of threshold					
One-factor model	1076	42502.433	0.791	0.835	0.040
Two-factor model	1074	24764.793	0.880	0.905	0.030
Three-factor model	1072	23971.704	0.884	0.908	0.030

Note: All chi-square different tests were statistically significant at 0.01 and were not listed in the table.

4.3 Invariance of Item Parameters

4.3.1 Generation of Item Parameters

A three-dimensional structure was shown to underlie the item responses to these 30 linking items. The item parameters for these items, modeled using a three-dimensional coordinate system, are given in Table 4.7. The lower asymptote ***c*** parameters in the first column were calibrated by BILOG-MG from the 72,277 examinees from three years of administrations. Also included were the vectors of **a** and **d** parameters calibrated by NOHARM under a three-dimensional compensatory logistic model. Items with the highest loadings were used to define the axes, which were rearranged for calibration with item 8 (L08) placed at the first place. Item 18 (G18) and item 23 (V23) followed as the second and the third items. By default, the *a*-parameters for the first item (L08) were set to 0 for a2 and a3 and 0 for the second item (G18) was set for a3.

These item parameters were fixed and input to TESTFACT for estimating the ability distribution for each year of administration. As shown in Table 4.8, the means were close to zero but there were differences in the mean levels in terms of the three constructs across years. The examinees who took the test in the third year had the highest mean ability levels (all equal to 0.1), while those in the sixth year were relatively low. Item responses for the test in the first year resulted in a correlation of 0.46 between θ_1 and θ_2, 0.42 between θ_1 and θ_3, and 0.64 between θ_2 and θ_3. For the third year, the correlations were 0.48 between θ_1 and θ_2, 0.40 between θ_1 and θ_3, and 0.64 between θ_2 and θ_3. The test data at the third year showed that θ_1 and θ_2 were correlated 0.53, θ_1 and θ_3 correlated 0.43, and θ_2 and θ_3 correlated 0.62. These correlations were corrected for attenuation using the formula given by:

$$r_{\theta x'\theta y'} = \frac{r_{\theta x\theta y}}{\sqrt{r_{\theta xx} r_{\theta yy}}}. \tag{4.1}$$

$r_{\theta x\theta y}$ was the observed correlations between θ estimates for each dimension. These values were estimated from the variables with measurement error. To model error-free measures of the constructs, the correlation matrix $r_{\theta x'\theta y'}$ was computed as a function of the reliabilities of the trait scores and should be higher than the observed correlations. The reliability values were obtained from the TESTFACT output as the average reliability measures over different proficiency levels for each dimension.

The means and variance-covariance matrices in Table 4.9 were used for generating item response data for each administration. The same set of item parameter estimates in Table 4.7 was input into the three-dimensional three-parameter logistic compensatory model for data generation. As a result, the simulated test data assumed the item parameter estimates were equivalent except for measurement error. On the other hand, samples were randomly selected from real data at each year of administration.

Table 4.8 Parameter estimates for linking items used for simulation

Item	c	a1	a2	A3	d
L01	0.267	**0.585**	0.032	0.041	1.270
L02	0.229	**0.575**	0.071	0.011	0.979
L03	0.344	**0.733**	0.005	0.064	-0.458
L04	0.079	**0.532**	0.129	0.037	0.697
L05	0.270	**0.879**	0.248	0.079	1.008
L06	0.125	**0.452**	0.086	0.081	0.538
L07	0.076	**0.603**	0.134	0.069	0.990
L08 *	0.384	**1.186**	0	0	1.009
L09	0.168	**0.802**	0.036	0.116	0.749
L10	0.000	**0.419**	0.087	0.132	0.229
G11	0.119	0.317	**0.567**	0.331	-0.121
G12	0.101	0.135	**0.398**	0.160	0.112
G13	0.063	0.087	**0.619**	0.185	0.501
G14	0.328	0.146	**0.456**	0.265	0.501
G15	0.140	0.405	**0.388**	0.100	1.542
G16	0.000	0.596	**0.745**	0.006	0.724
G17	0.500	0.350	0.524	**0.784**	-0.103
G18 *	0.023	0.030	**0.971**	0	0.590
G19	0.284	**0.591**	0.466	0.292	-0.254
G20	0.190	0.234	**0.317**	0.224	0.514
V21	0.200	0.468	0.152	**0.908**	-0.351
V22	0.500	0.437	0.441	**0.857**	0.862
V23 *	0.200	0.026	0.209	**1.028**	-1.055
V24	0.315	0.283	0.357	**0.489**	0.078
V25	0.255	0.062	**0.555**	0.520	-0.747
V26	0.121	0.123	**0.518**	0.239	0.213
V27	0.125	0.167	**0.454**	0.083	0.180
V28	0.071	0.258	**0.669**	0.246	0.291
V29	0.015	0.109	**0.681**	0.190	0.405
V30	0.016	0.294	**0.362**	0.250	-0.626

Note: * Items with the highest loadings were used to anchor the axes.

Table 4.9 Estimates for ability distribution used for simulation

	Year 1			Year 3			Year 6		
Descriptives	θ1	θ2	θ3	θ1	θ2	θ3	θ1	θ2	θ3
Mean									
	0	-0.1	0.1	0.1	0.1	0.1	-0.1	0	-0.1
Variance/Covariance									
θ1	0.6	0.2	0.1	0.6	0.1	0.1	0.6	0.2	0.2
θ2	0.2	0.7	0.2	0.1	0.6	0.2	0.2	0.6	0.3
θ3	0.1	0.2	0.4	0.1	0.2	0.4	0.2	0.3	0.4

4.3.2 Results for Model I

The plots summarizing the means of the item parameter estimates were displayed for Model I over years in Figures 4.5 and 4.6. This model assumed a dominant one-factor solution for all 30 items. These item parameter estimates were linked to the same reference scale calibrated on all the data with three administrations combined.

Figure 4.5 compared estimates for *a* parameters and Figure 4.6 contrasted estimates for *b* parameters. The panels in the top showed the means for parameters estimates from 400 simulation samples and the bottom one exhibited the means for parameters estimates from 400 samples from the real data. As expected, the means of simulation samples were fairly similar across years of administration and nearly fell on the same line, which suggested the items had invariant difficulty values and discriminate equally well over time. A few exceptions were item 17 and item 22 in terms of the item discrimination.

Compared to simulation data, the means for real samples almost fell on a line but have a clear variation across administration. Estimates of most item difficulty parameters remained stable over years except for item 3 and item 29. The item discrimination power were relatively variable, especially for item 5, item 17, item 21, and item 22.

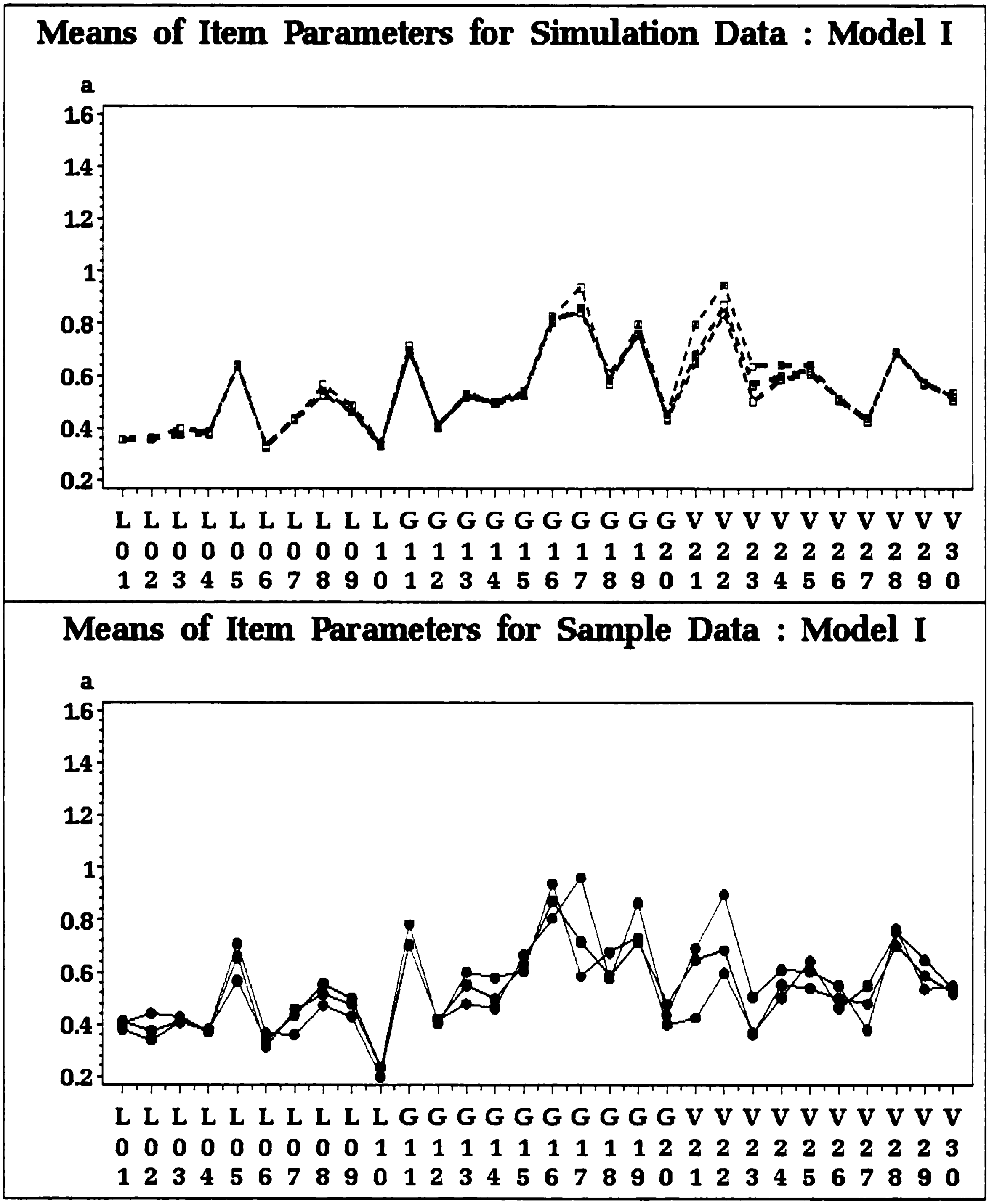
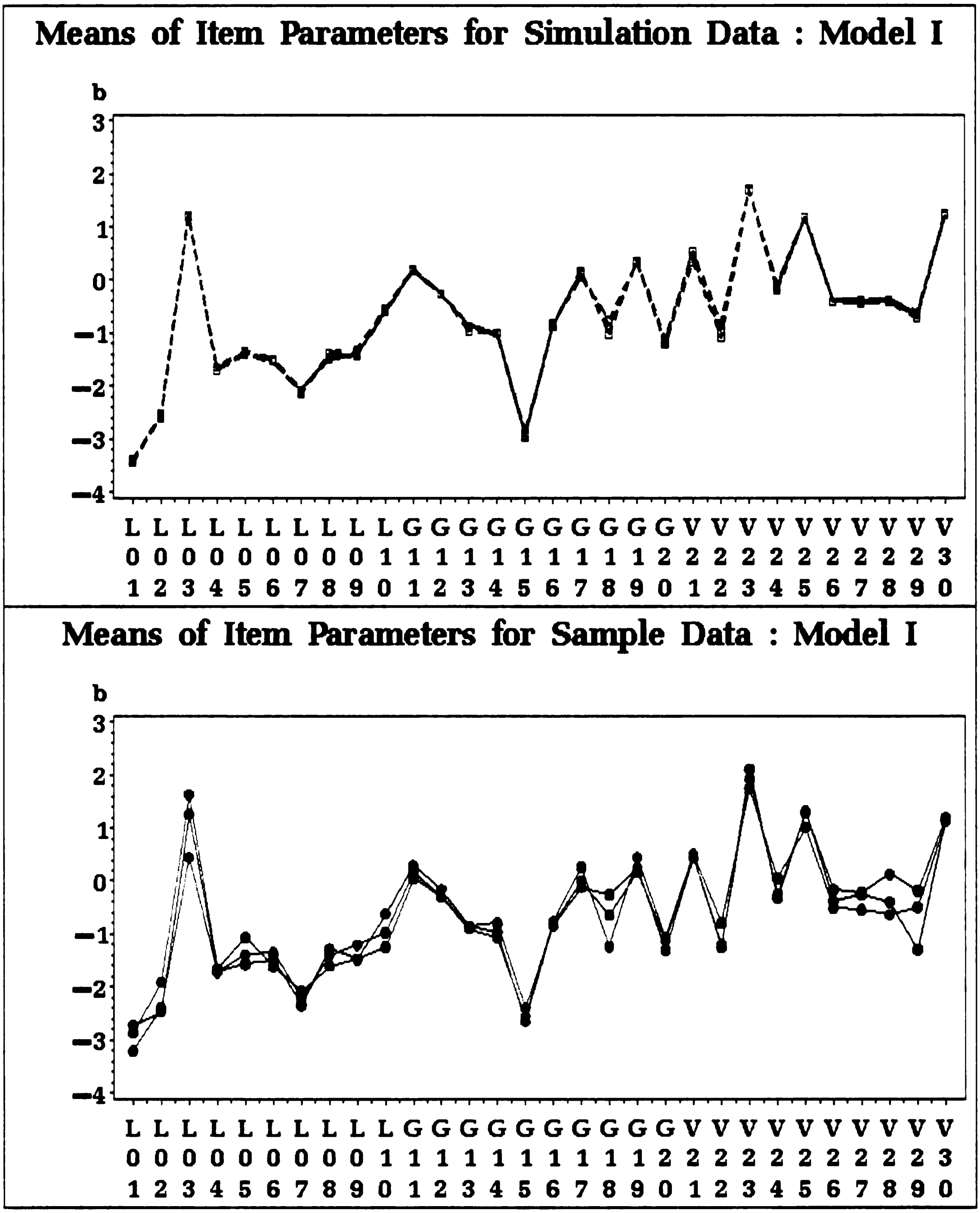


Figure 4.5 Mean plots of estimates for *a* parameters for simulation data and real data underlying Model I[1]
(Images in this dissertation are presented in color)

[1] The dotted lines represent the parameter estimates based on simulation sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.


Means of Item Parameters for Simulation Data : Model I

Means of Item Parameters for Sample Data : Model I

Figure 4.6 Mean plots of estimates for *b* parameters for simulation data and real data underlying Model I[2]
(Images in this dissertation are presented in color)

[2] The dotted lines represent the parameter estimates based on simulation sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

Multivariate analysis of variance was performed to test the differences in the mean vectors for both parameter estimates in terms of the design (simulation versus real) and the year of administration (year 1, year 3, and year 6). Results of multivariate tests of group differences are listed in Table 4.10.

Table 4. 10 Results of MANOVA test criteria, F-statistics, and η^2 for Model I

	Overall						a		b	
Item	Design			Year			Design	Year	Design	Year
	Λ	$F_{2,2393}$	η^2	Λ	$F_{4,4786}$	η^2	$F_{1,2394}$	$F_{2,2394}$	$F_{1,2394}$	$F_{2,2394}$
L01	0.81	274.5	0.19	0.81	129.8	0.10	343.6	16.6	527.6	36.4
L02	0.78	343.6	0.22	0.30	992.9	0.45	184.4	194.4	512.1	195.0
L03	0.92	98.1	0.08	0.32	916.1	0.43	155.4	38.8	162.7	1826.8
L04	1.00	5.1*	0.00	0.93	42.6	0.03	8.6*	2.1*	3.4*	22.1
L05	0.98	21.1	0.02	0.41	669.1	0.36	5.2*	241.8	33.7	758.2
L06	0.98	30.5	0.02	0.89	69.3	0.05	36.1	114.2	5.1*	136.2
L07	0.95	66.6	0.05	0.53	445.5	0.27	63.7	254.3	124.8	87.8
L08	0.87	182.1	0.13	0.68	256.8	0.18	137.5	24.7	0.9*	222.9
L09	1.00	2.4*	0.00	0.47	546.5	0.31	0.6*	37.6	3.8*	419.5
L10	0.29	2975.5	0.71	0.52	470.6	0.28	5838.5	30.0	2864.5	694.5
G11	0.90	125.7	0.10	0.52	456.0	0.28	123.9	82.0	114.9	971.1
G12	1.00	4.2	0.00	0.86	94.6	0.07	8.3	21.5	0.6*	149.1
G13	0.95	64.0	0.05	0.74	190.1	0.14	91.4	382.4	121.0	112.4
G14	0.95	60.2	0.05	0.84	112.3	0.09	54.8	208.1	119.2	201.2
G15	0.67	583.0	0.33	0.71	225.4	0.16	1166.3	48.3	943.2	39.9
G16	0.86	191.3	0.14	0.73	201.8	0.14	382.5	175.5	154.5	17.2
G17	0.69	530.6	0.31	0.34	860.8	0.42	799.6	982.8	533.1	1911.4
G18	0.45	1463.1	0.55	0.07	3259.3	0.73	110.5	410.3	2297.3	10693.4
G19	0.80	297.3	0.20	0.42	659.1	0.36	0.1*	319.5	594.8	1330.2
G20	0.99	8.0	0.01	0.68	250.2	0.17	6.9*	183.7	0.1*	321.4
V21	0.57	915.7	0.43	0.28	1082.3	0.47	1811.8	1822.4	22.8	673.7
V22	0.61	769.6	0.39	0.43	632.5	0.35	1458.3	871.0	444.5	1436.3
V23	0.43	1617.8	0.57	0.46	566.9	0.32	2532.2	722.6	550.3	123.4
V24	0.83	247.7	0.17	0.36	784.2	0.40	419.3	316.7	249.1	1931.1
V25	0.97	37.0	0.03	0.64	302.7	0.20	67.4	87.4	11.8	425.9
V26	0.88	159.6	0.12	0.37	761.7	0.39	10.2	204.0	203.7	997.1
V27	0.82	256.1	0.18	0.48	525.9	0.31	495.0	746.7	213.3	995.8
V28	0.63	712.1	0.37	0.10	2610.8	0.69	444.8	62.9	1399.0	8319.6
V29	0.96	45.9	0.04	0.06	3719.9	0.76	75.8	283.2	2.4*	11080.9
V30	0.85	210.4	0.15	0.94	36.1	0.03	66.6	51.4	353.8	21.2

Note: * These F statistics were statistically not significant at p value of 0.01 and univariate tests for *a* and *b* used Bonferroni correction at $p \geq 0.005$.

The hypothesis assuming mean vectors for estimates were identical was rejected between simulation data and real data except for two items. Also rejected was the null hypothesis that the mean vectors remained invariant across three administrations. At least two years differed in means for *a* and *b*. Though the difference between mean vectors of both item difficulty and item discrimination were statistically significant, the effect sizes were relatively small. The patterns between simulation data and real data were similar. There were no substantial differences between item parameter estimates.

The Bonferroni corrected ANOVA was included in Table 4.10 for each individual item parameter. Each test was carried out with 1 and 2,394 degree of freedom for design and 2 and 2,394 degree of freedom for year. With only two dependent variables, a 0.01 level of test would be rejected if the *p*-values were less than 0.005 under a Bonferroni correction. It appeared that both variables showed a significant year effect for almost all items. However, there was a comparatively weaker effect of design for five or six items.

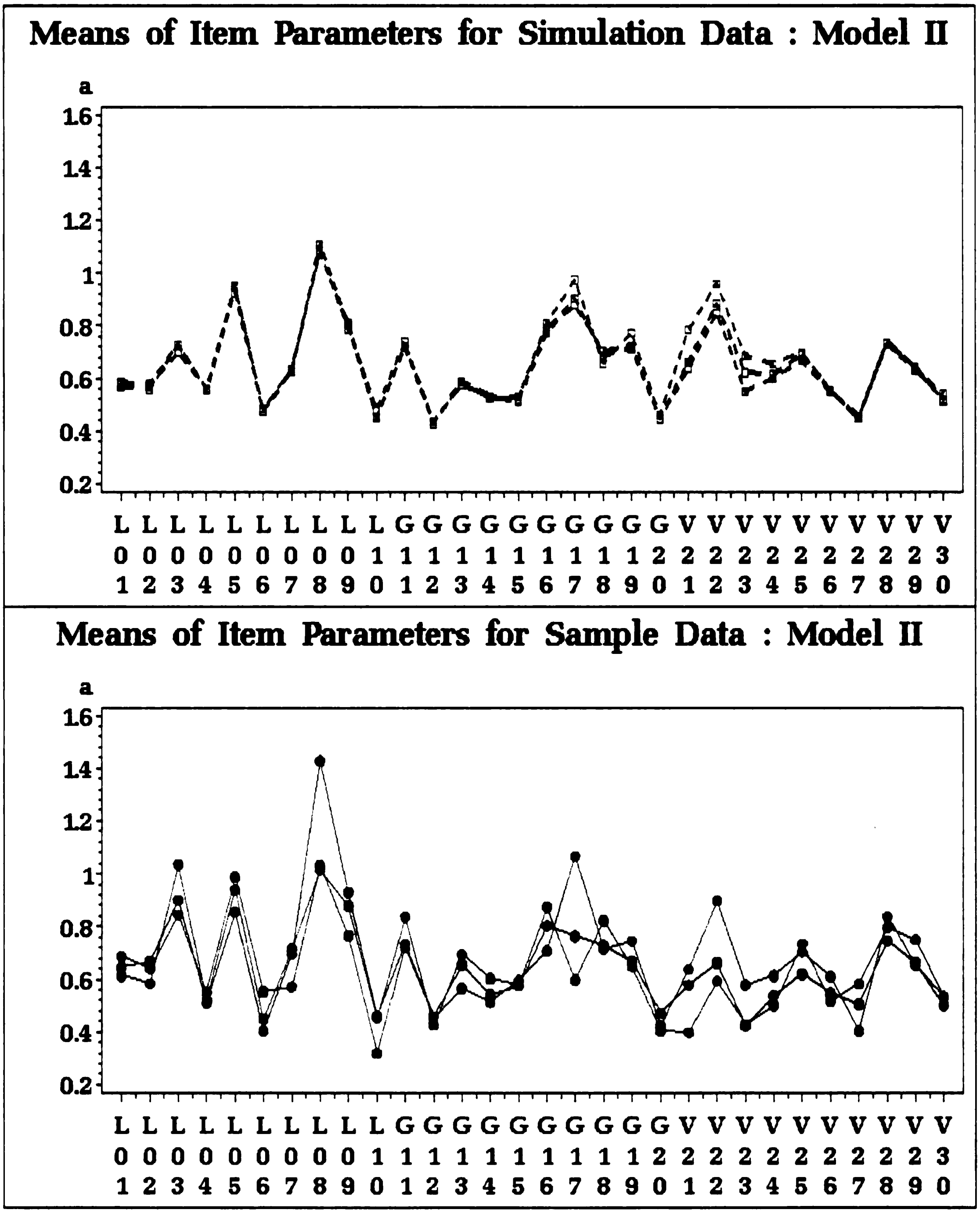
Even though most of the *F*-statistics were statistically significant, the magnitude of effect should also be taken into account. The eta-squares (η^2) were computed to measure the size of overall effect for design and years of administration. The values had a wide range from zero to 0.71 for design effect and the maximum is 0.76 for year effect. However, in general the eta-squares were moderate for year effect but relatively low for design effect, which suggested the mean vectors of both parameter estimates had a significant variation over years while the differences in mean vectors were negligible for some items between simulation and real data.

4.3.3 Results for Model II

The plots summarizing the means of the item parameter estimates are displayed for Model II over years in Figures 4.7 and 4.8. This model assumed two different test scales for all 30 items with one indicating oracy skills and the other indicating literacy skills. These item parameter estimates were linked to the same scale. The reference scale was the parameter estimates calibrated on all the data with three administrations combined.

Figure 4.7 compared estimates for *a* parameters and Figure 4.8 contrasts estimates for *b* parameter. The panels on the top showed the means for parameters estimates from 400 simulation samples and the bottom one exhibited the means for parameters estimates from 400 samples from the real data. As expected, the means of simulation samples were fairly similar across years of administration and nearly fall on the same line, which suggested the items have the invariant difficulty values and discriminate equally well over time. A few exceptions were item 17 and item 22 in terms of the item discrimination.

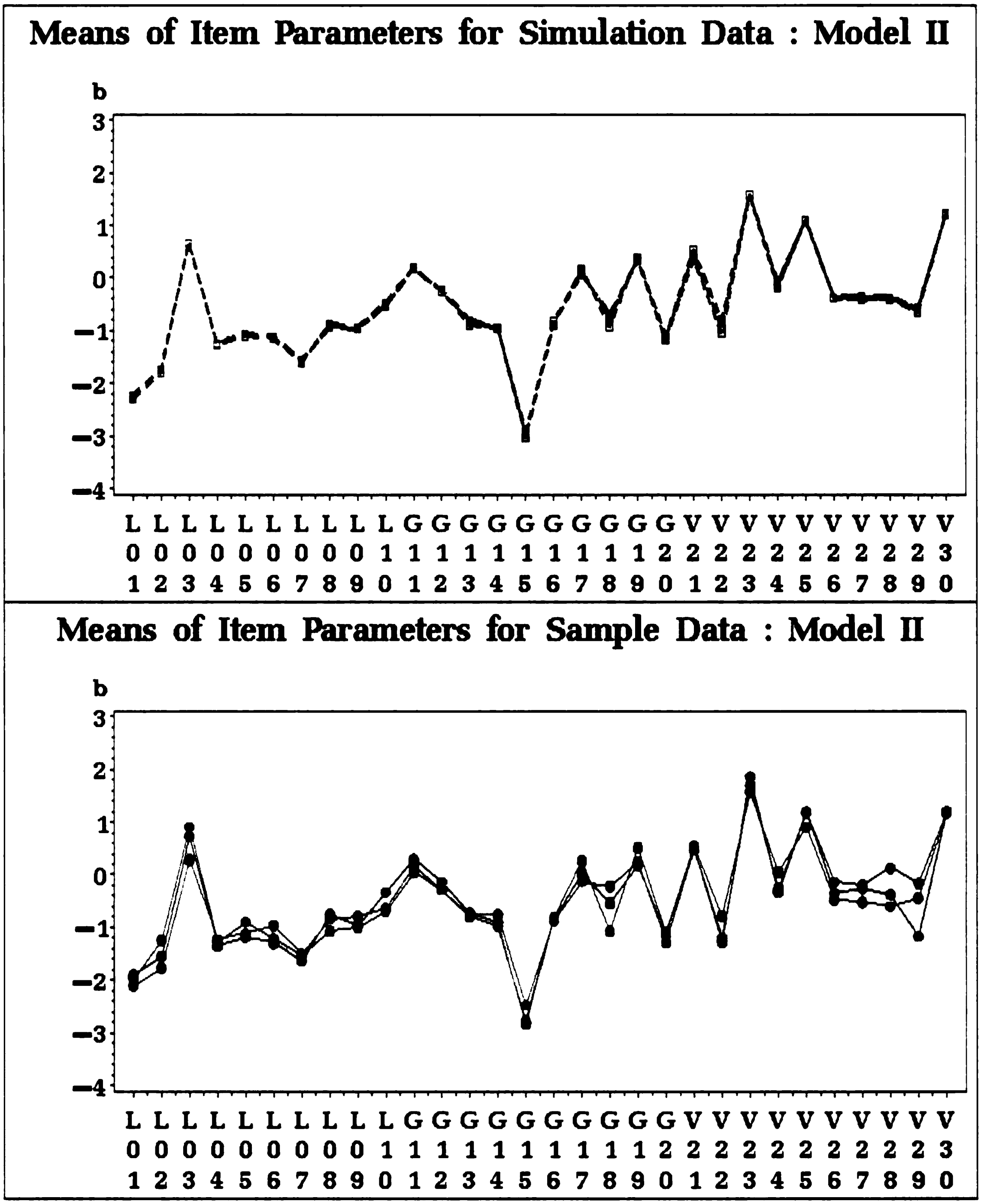
Compared to the simulated data, the means for real samples almost fell on a line but had clear variation across administration. Estimates of most item difficulties remained stable over years except for item 3 and item 29. The item discrimination power was relatively variable, especially for that of item 3, item 17, item 21, and item 22.


Means of Item Parameters for Simulation Data : Model II

Means of Item Parameters for Sample Data : Model II

Figure 4.7 Mean plots of estimates for *a* parameter for simulation data and real data underlying Model II[3]
(Images in this dissertation are presented in color)

[3] The dotted lines represent the parameter estimates based on simulation sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.


Means of Item Parameters for Simulation Data : Model II

Means of Item Parameters for Sample Data : Model II

Figure 4.8 Mean plots of estimates for *b* parameter for simulation data and real data underlying Model II[4]
(Images in this dissertation are presented in color)

[4] The dotted lines represent the parameter estimates based on simulation sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

Multivariate analysis of variance was performed to test the differences in the mean vectors for both parameter estimates in terms of the design (simulation versus real) and the year of administration (year 1, year 3, and year 6). Results of multivariate tests of group differences were listed in Table 4.11.

Table 4.11 Results of MANOVA test criteria, F-statistics, and η^2 for Model II

Item	Overall						a		b	
	Design			Year			Design	Year	Design	Year
	Λ	$F_{2,2393}$	η^2	Λ	$F_{,4786}$	η^2	$F_{1,2394}$	$F_{2,2394}$	$F_{1,2394}$	$F_{2,2394}$
L01	0.81	282.9	0.19	0.72	214.4	0.15	424.0	28.4	565.1	45.0
L02	0.72	469.2	0.28	0.32	904.5	0.43	372.3	63.9	823.0	411.2
L03	0.69	531.6	0.31	0.30	976.4	0.45	939.9	47.6	60.7	1566.9
L04	0.92	98.6	0.08	0.91	55.8	0.04	133.2	29.5	34.1	70.1
L05	0.98	22.6	0.02	0.52	462.7	0.28	8.4*	78.2	2.4*	509.1
L06	1.00	4.3*	0.00	0.75	189.7	0.14	8.1*	374.8	7.4*	374.8
L07	0.94	82.3	0.06	0.49	511.1	0.30	83.1	185.0	9.1	67.5
L08	0.96	45.3	0.04	0.51	477.0	0.29	89.0	375.2	34.1	661.4
L09	0.92	98.8	0.08	0.44	607.9	0.34	175.2	124.8	40.6	356.0
L10	0.85	203.0	0.15	0.49	507.9	0.30	405.3	229.7	103.7	614.7
G11	0.91	117.7	0.09	0.50	492.1	0.29	170.0	135.6	53.6	996.3
G12	0.97	32.9	0.03	0.82	126.7	0.10	65.6	61.4	7.4*	204.7
G13	0.78	346.2	0.22	0.74	193.8	0.14	678.9	374.7	415.8	167.0
G14	0.93	87.8	0.07	0.89	71.3	0.06	98.1	97.8	175.7	144.5
G15	0.84	235.6	0.16	0.72	211.5	0.15	467.1	16.3	418.0	82.4
G16	0.99	6.7	0.01	0.66	277.6	0.19	4.4*	257.0	13.2	27.1
G17	0.76	381.4	0.24	0.32	914.9	0.43	491.8	1171.2	476.0	1981.8
G18	0.42	1672.0	0.58	0.07	3221.3	0.73	851.9	313.8	3274.1	10575.8
G19	0.84	229.2	0.16	0.36	791.7	0.40	234.3	216.1	205.9	1826.2
G20	0.96	54.9	0.04	0.70	232.1	0.16	87.5	112.8	18.3	253.0
V21	0.44	1495.7	0.56	0.28	1082.3	0.47	2991.7	1571.7	233.3	767.4
V22	0.58	868.6	0.42	0.42	649.9	0.35	1693.9	841.6	599.9	1502.2
V23	0.48	1301.8	0.52	0.45	584.7	0.33	1725.4	652.9	218.8	103.0
V24	0.76	382.5	0.24	0.36	788.4	0.40	704.6	331.6	285.7	1989.5
V25	1.00	4.6*	0.00	0.62	318.3	0.21	0.1*	69.7	5.6*	483.3
V26	0.89	153.3	0.11	0.37	781.3	0.40	7.9	195.9	278.9	1050.7
V27	0.81	279.9	0.19	0.48	523.5	0.30	537.7	703.4	245.7	1001.9
V28	0.61	760.5	0.39	0.10	2668.3	0.69	493.0	92.7	1490.8	8157.7
V29	0.80	295.9	0.20	0.07	3271.5	0.73	524.2	189.5	35.4	10402.2
V30	0.93	90.9	0.07	0.93	46.7	0.04	0.2*	68.8	75.4	13.3

Note: * These F statistics were statistically not significant at p value of 0.01 and univariate tests for *a* and *b* used Bonferroni correction at $p \geq 0.005$.

The hypothesis assuming mean vectors for estimates were identical was rejected between simulation data and real data except for two items. Also rejected was the null hypothesis that the mean vectors remained invariant across three administrations. At least two years differed in means for *a* and *b*. Though the difference between mean vectors of both item difficulty and item discrimination were statistically significant, the effect sizes were relatively small. The patterns between simulation data and real data were similar. There were no substantial differences between item parameter estimates.

The Bonferroni corrected ANOVA were included in Table 4.11 for each individual item parameter. Each test was carried out with 1 and 2,394 degree of freedom for design and 2 and 2,394 degree of freedom for year. With only two dependent variables, a 0.01 level of test would be rejected if the *p*-values were less than 0.005 under a Bonferroni correction. It appeared that both variables showed a significant year effect for all items. However, there was a comparatively weak effect of design for five or six items.

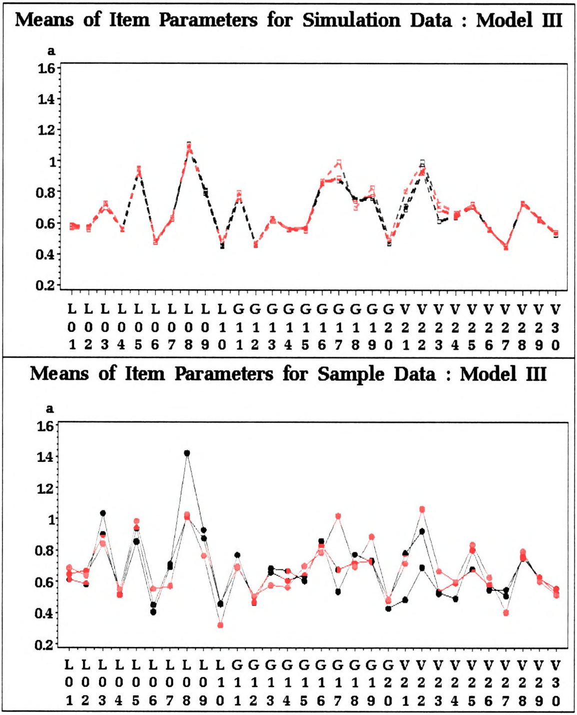
Even though most of the *F*-statistics were statistically significant, the magnitude of effect should also be taken into account. The eta-squares (η^2) were computed to measure the size of overall effect for design and years of administration. The values had a wide range from 0 to 0.58 for design effect and the maximum is 0.73 for year effect. However, in general the eta-squares were moderate for year effect but relatively low for design effect, which suggested the mean vectors of both parameter estimates had a significant variation over years while the differences in mean vectors were negligible for some item between simulation and real data.

4.3.4 Results for Model III

The plots summarizing the means of the item parameter estimates were displayed for Model III over years in Figures 4.9 and 4.10. This model assumed each set of items in sections of listening, grammar, and vocabulary represent a test subscale. These item parameter estimates were linked to the same scale. The reference scale was the parameter estimates calibrated on all the data with three administrations combined.

Figure 4.9 compared estimates for *a*-parameters and Figure 4.10 contrasts estimates for *b*-parameters. The panels on the top showed the means for parameter estimates from 400 simulation samples and the bottom one exhibits the means for parameter estimates from 400 samples from the real data. As expected, the means of simulation samples were fairly similar across years of administration and nearly fell on the same line, which suggested the items have the invariant difficulty values and discriminate equally well over time. An exception was item 17 in terms of item discrimination.

Compared to simulation data, the means for real samples almost fell on a line but had clear variation across administration. Estimates of most item difficulties remained stable over years except for that of item 18 and item 29. The item discrimination power was relatively variable, especially for that of item 8, item 17, item 21, and item 22.



Means of Item Parameters for Simulation Data : Model III

Means of Item Parameters for Sample Data : Model III

Figure 4.9 Mean plots of estimates for *a* parameter for simulation data and real data assuming Model III[5]
(Images in this dissertation are presented in color)

[5] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

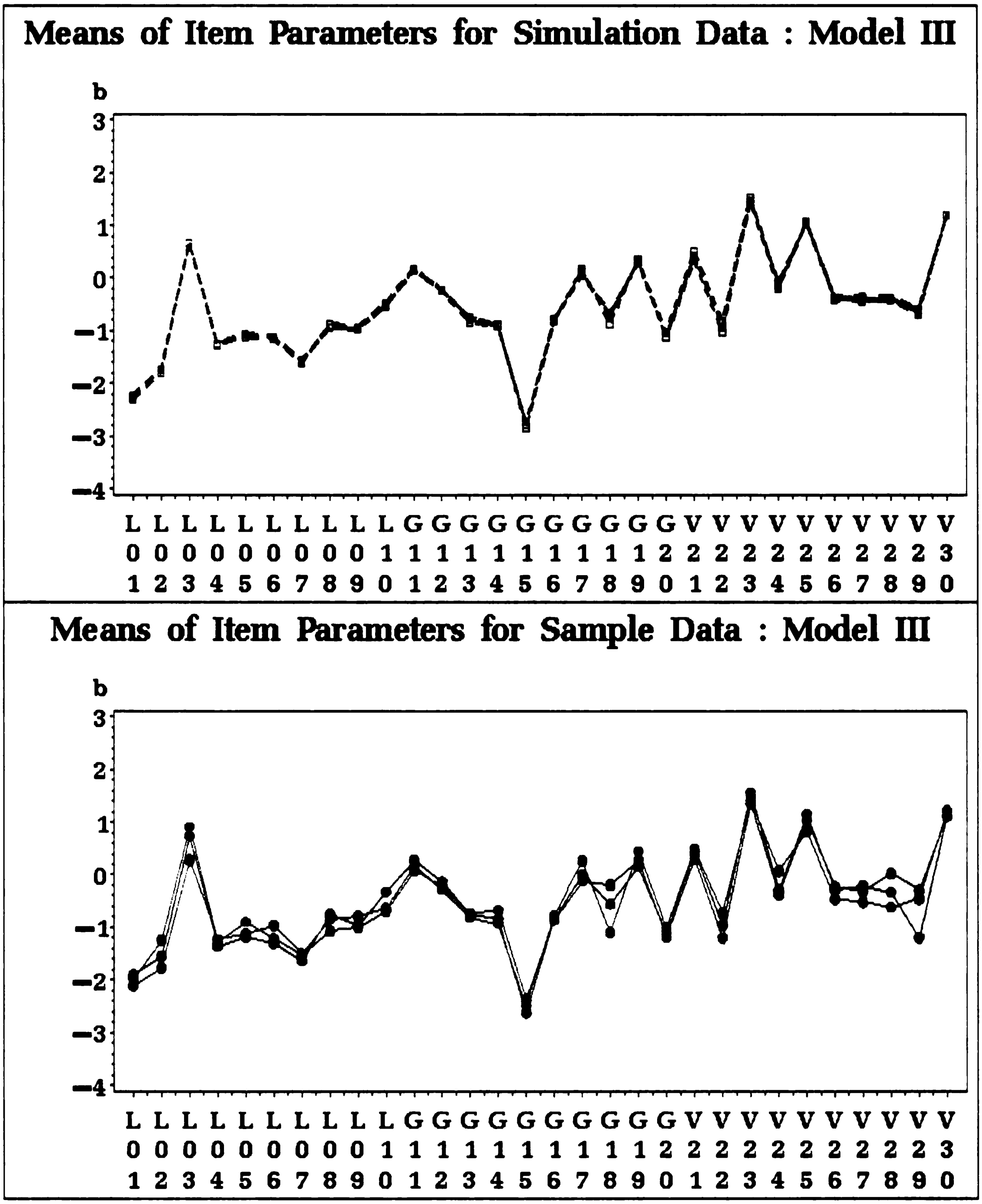


Figure 4.10 Mean plots of estimates for *b* parameter for simulation data and real data assuming Model III[6]
(Images in this dissertation are presented in color)

[6] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

Multivariate analysis of variance was performed to test the differences in the mean vectors for both parameter estimates in terms of the design (simulation versus real) and the year of administration (year 1, year 3, and year 6). Results of multivariate tests of group differences are listed in Table 4.12.

Table 4.12 Results of MANOVA test criteria, F-statistics, and η^2 for Model III

Item	Overall						a		b	
	Design			Year			Design	Year	Design	Year
	Λ	$F_{2,2393}$	η^2	Λ	$F_{,4786}$	η^2	$F_{1,2394}$	$F_{2,2394}$	$F_{1,2394}$	$F_{2,2394}$
L01	0.81	283.1	0.19	0.72	214.0	0.15	425.9	28.3	565.6	44.8
L02	0.72	471.9	0.28	0.32	909.2	0.43	373.0	63.6	826.2	413.5
L03	0.69	535.1	0.31	0.30	980.8	0.45	946.3	49.1	61.2	1571.8
L04	0.92	99.4	0.08	0.91	56.2	0.04	134.2	29.6	34.4	70.5
L05	0.98	21.9	0.02	0.52	465.2	0.28	7.9*	78.8	2.4*	511.8
L06	1.00	4.4*	0.00	0.74	190.0	0.14	8.4*	375.2	7.6*	375.8
L07	0.94	82.9	0.06	0.49	513.0	0.30	83.3	185.2	9.0*	68.1
L08	0.96	46.8	0.04	0.51	476.6	0.28	92.0	373.2	34.9	659.9
L09	0.92	99.4	0.08	0.44	609.5	0.34	175.9	124.5	40.3	357.6
L10	0.85	205.3	0.15	0.49	508.7	0.30	409.9	229.8	104.7	615.5
G11	0.90	134.8	0.10	0.59	360.1	0.23	269.7	38.3	5.1*	819.6
G12	0.91	116.4	0.09	0.77	162.7	0.12	232.8	59.9	48.2	221.4
G13	0.98	26.5	0.02	0.83	113.8	0.09	43.5	212.1	47.8	191.7
G14	0.87	179.8	0.13	0.85	102.4	0.08	330.0	92.3	308.4	206.8
G15	0.79	327.6	0.21	0.65	284.3	0.19	628.7	83.1	409.2	54.2
G16	0.95	66.9	0.05	0.84	110.6	0.08	112.8	31.5	120.0	43.0
G17	0.69	527.3	0.31	0.33	901.4	0.43	845.2	909.1	340.0	1884.7
G18	0.49	1242.1	0.51	0.07	3284.4	0.73	0.1*	136.1	1414.2	10326.4
G19	0.92	98.7	0.08	0.39	723.1	0.38	1.1*	273.2	183.8	1300.4
G20	0.98	28.9	0.02	0.77	171.3	0.13	50.6	115.8	17.6	268.2
V21	0.83	249.3	0.17	0.27	1091.3	0.48	374.7	876.2	29.9	1386.2
V22	0.95	62.4	0.05	0.48	535.7	0.31	110.5	468.2	101.5	1257.2
V23	0.49	1225.3	0.51	0.43	629.8	0.34	572.9	327.2	55.3	105.9
V24	0.75	400.7	0.25	0.35	822.1	0.41	674.8	156.5	369.7	2052.4
V25	0.93	90.9	0.07	0.75	181.3	0.13	153.7	86.6	150.6	334.2
V26	0.92	101.2	0.08	0.54	425.9	0.26	105.2	80.8	179.2	543.5
V27	0.88	157.8	0.12	0.57	384.1	0.24	313.1	366.4	110.5	684.8
V28	0.78	331.4	0.22	0.17	1707.3	0.59	143.1	10.4	662.7	4256.5
V29	0.98	23.8	0.02	0.12	2196.2	0.65	11.9	13.8	46.7	5111.9
V30	0.92	107.4	0.08	0.87	82.7	0.06	0.1*	23.0	91.6	58.4

Note: * These F statistics were statistically not significant at p value of 0.01 and univariate tests for *a* and *b* used Bonferroni correction at $p \geq 0.005$.

The hypothesis assuming mean vectors for estimates were identical was rejected between simulation data and real data except for two items. Also rejected was the null hypothesis that the mean vectors remained invariant across three administrations. At least two years differed in means for *a* and *b*. Though the difference between mean vectors of both item difficulty and item discrimination were statistically significant, the effect sizes were relatively small. The patterns between simulation data and real data were similar. There were no substantial differences between item parameter estimates.

The Bonferroni corrected ANOVA are included in Table 4.12 for each individual item parameter. Each test was carried out with 1 and 2,394 degree of freedom for design and 2 and 2,394 degree of freedom for year. With only two dependent variables, a 0.01 level of test would be rejected if the *p*-values were less than 0.005 under a Bonferroni correction. It appeared that both variables showed a significant year effect for all items. However, there was a comparatively weaker effect of design for five or six items.

Even though most of the *F*-statistics were statistically significant, the magnitude of effect should also be taken into account. The eta-squares (η^2) were computed to measure the size of overall effect for design and years of administration. The values had a wide range from 0 to 0.51 for design effect and the maximum is 0.87 for year effect. However, in general the eta-squares were moderate for year effect but relatively low for design effect, which suggested the mean vectors of both parameter estimates had significant variation over years while the differences in mean vectors were negligible for some item between simulation and real data.

4.3.5 Results for Model IV

The plots summarizing the means of the item parameter estimates are displayed for Model IV over years in Figures 4.11, 4.12, 4.13 and 4.14. This model assumed a three-dimensional extension of three-parameter compensatory models resulting in three skill areas. These item parameter estimates were linked to the same scale. The reference scale was the parameter estimates calibrated on all the data with three administrations combined.

Figures 4.11 compared estimates for a_1 parameters, and Figures 4.12 and 4.13 compare a_2 and a_3 parameter estimates respectively. Figure 4.14 contrasted estimates for d parameter. The panels on the top showed the means for parameters estimates from 400 simulation samples and the bottom one exhibited the means for parameters estimates from 400 samples from the real data. As expected, the means were approximately equivalent across years of administration and overlap on the same line, which suggested the items had invariant difficulty values and discriminated equally well over time. One exception was item 18 in terms of the a_2 parameter.

Compared to simulated data, the means for real samples almost fell on a line but had slight variation across administrations. Estimates of most item difficulties remained stable over years except for that of item 17. The item discrimination powers were relatively variable, especially for item 3 and item 8.

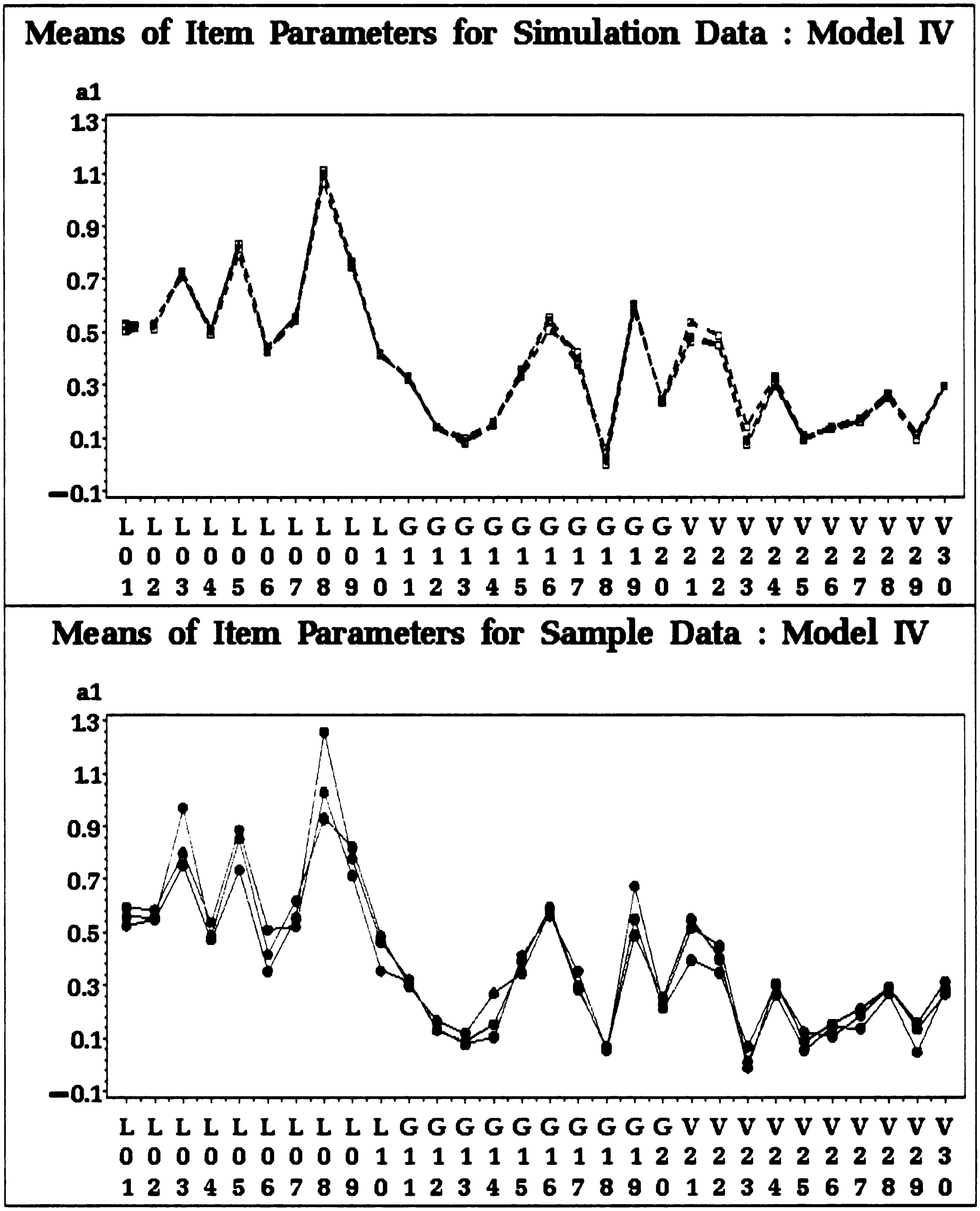
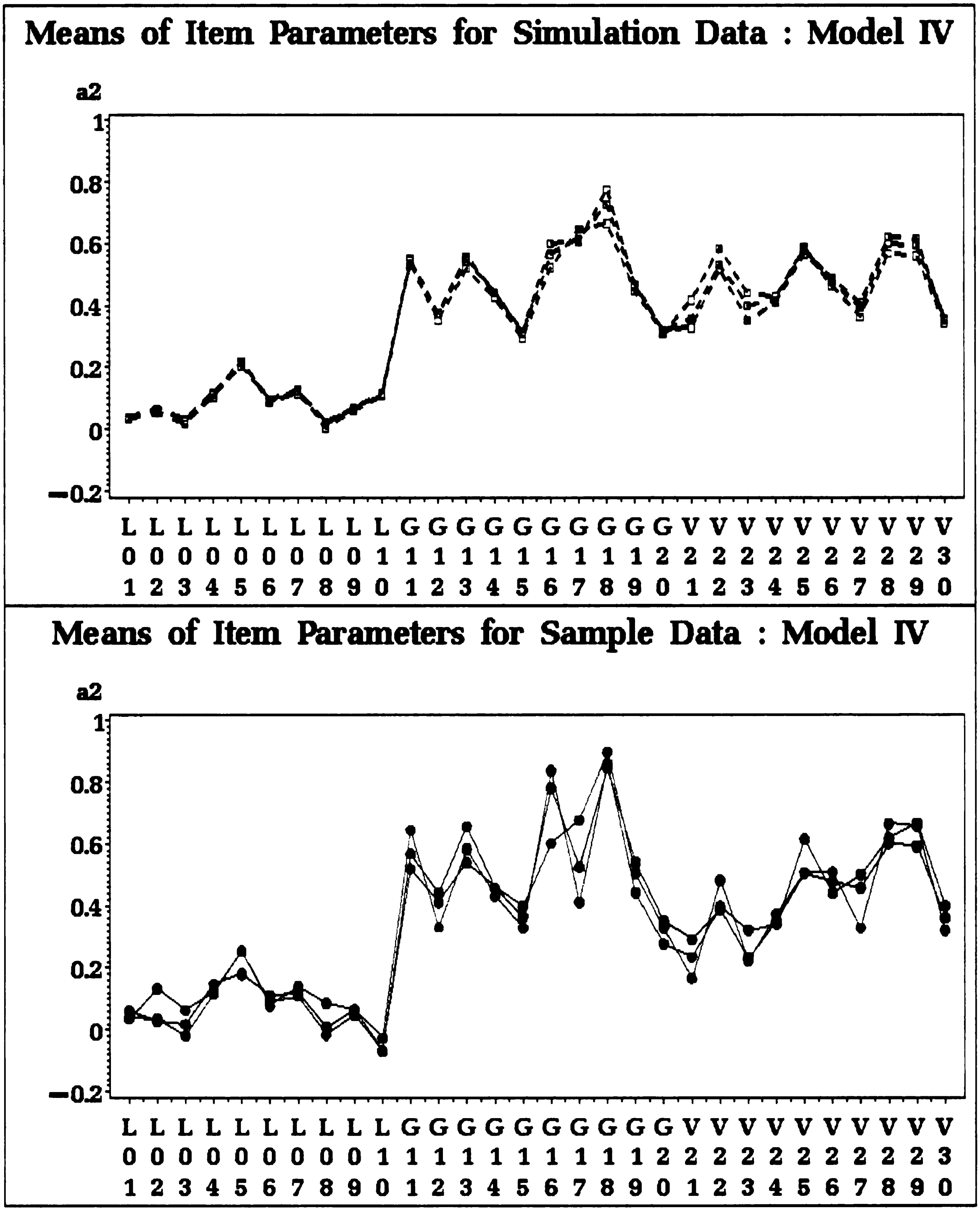


Figure 4.11 Mean plots of estimates for a_1 parameter for simulation data and real data underlying Model IV[7]
(Images in this dissertation are presented in color)

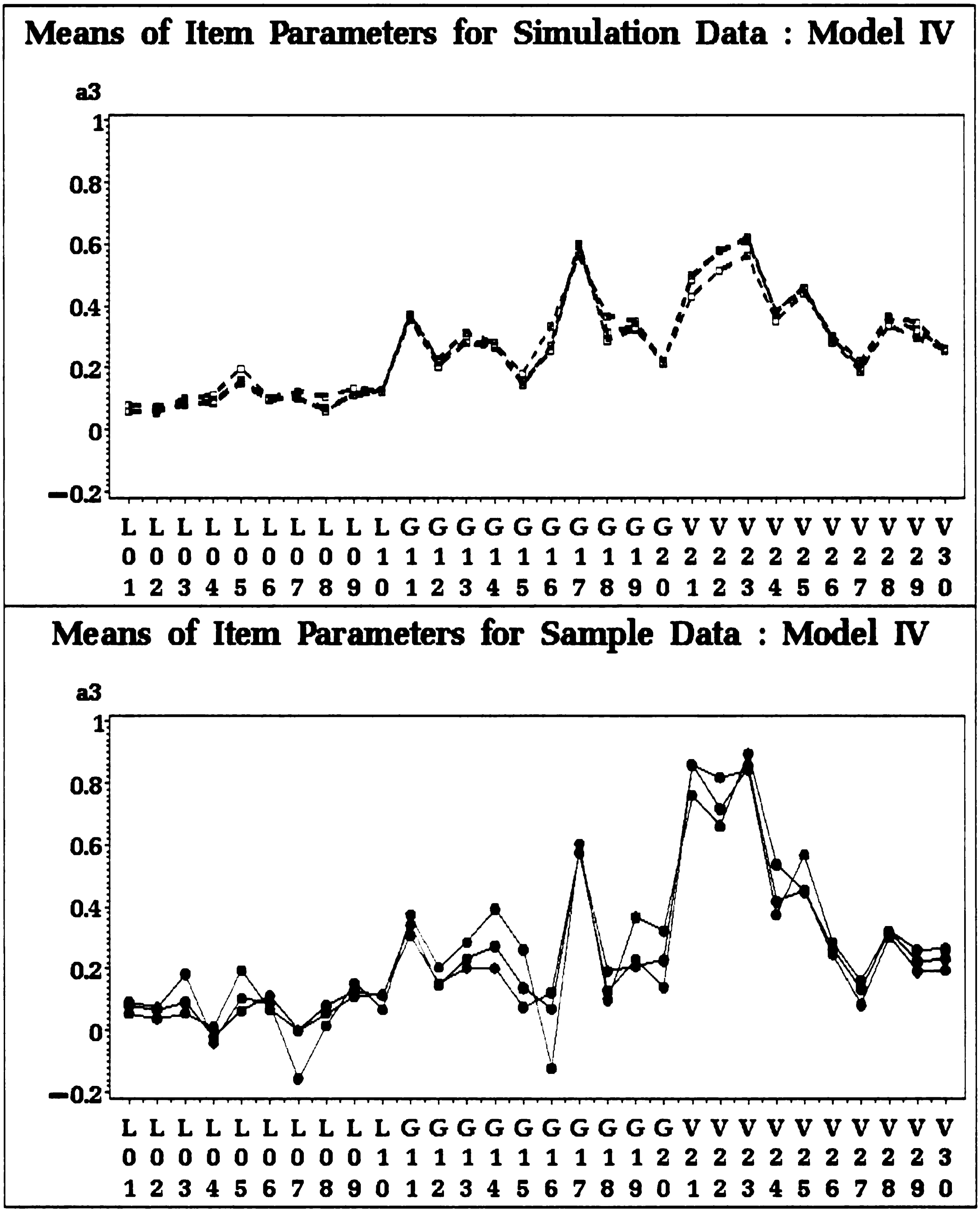
[7] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.


Means of Item Parameters for Simulation Data : Model IV

Means of Item Parameters for Sample Data : Model IV

Figure 4.12 Mean plots of estimates for a_2 parameter for simulation data and real data underlying Model IV[8]
(Images in this dissertation are presented in color)

[8] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.



Means of Item Parameters for Simulation Data : Model IV

Means of Item Parameters for Sample Data : Model IV

Figure 4.13 Mean plots of estimates for a_3 parameter for simulation data and real data underlying Model IV[9]
(Images in this dissertation are presented in color)

[9] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

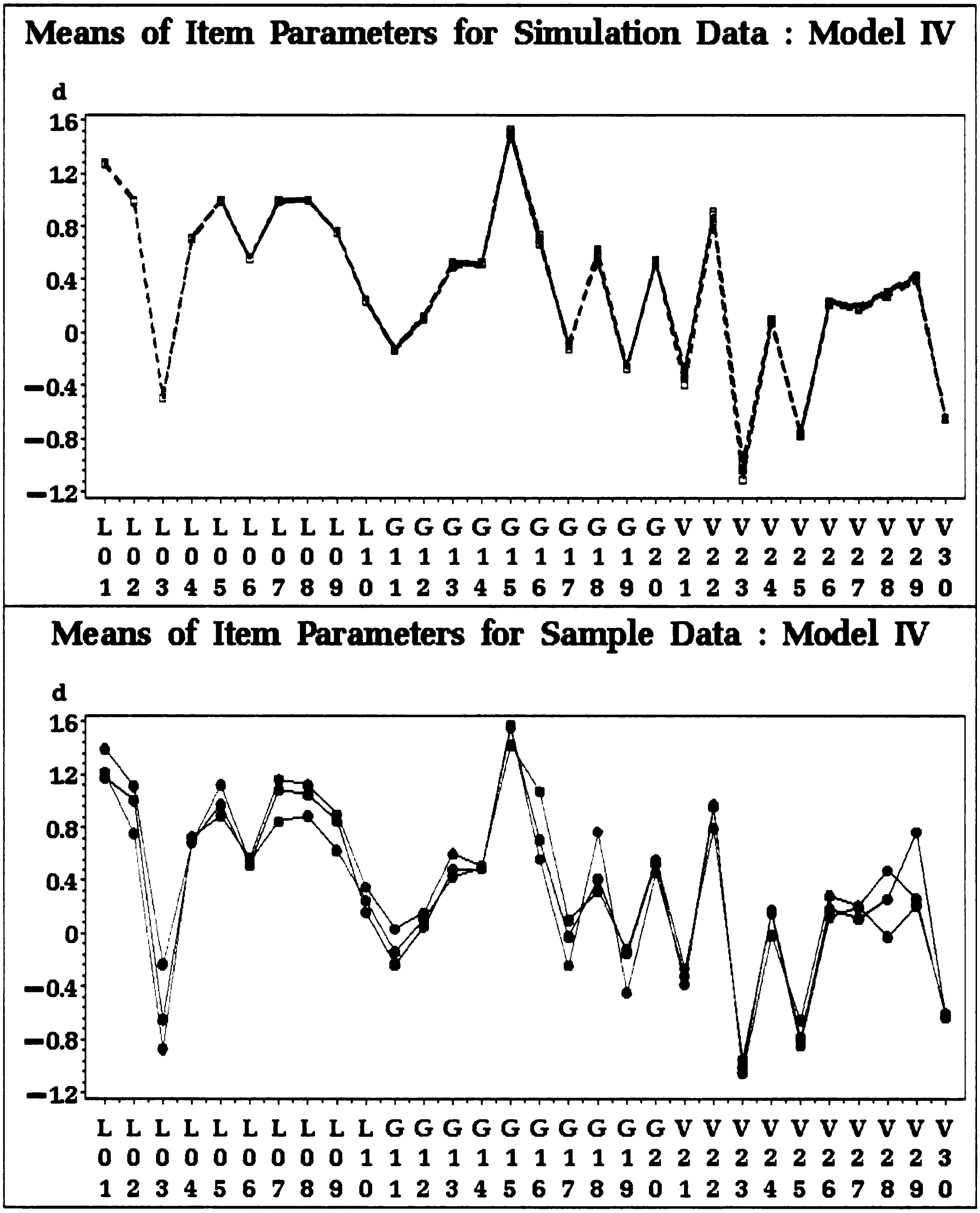


Figure 4.14 Mean plots of estimates for *d* parameter for simulation data and real data underlying Model IV[10]
(Images in this dissertation are presented in color)

[10] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

Multivariate analysis of variance was performed to test the differences in the mean vectors for all four parameter estimates in terms of the design (simulation versus real) and the year of administration (year 1, year 3, and year 6). Results of multivariate tests of group differences are listed in Tables 4.13 and 4.14.

The hypothesis assuming mean vectors for estimates were identical was rejected between simulation data and real data except for two items. Also rejected was the null hypothesis that the mean vectors remained invariant across three administrations, although two years differed in means for a_1, a_2, a_3, and d. Though the differences between mean vectors of both item difficulty and item discrimination were statistically significant, the patterns between simulation data and real data were similar. There were no substantial differences among item parameter estimates.

The Bonferroni corrected ANOVA were included in Table 4.13 for each individual item parameter. With only two dependent variables, a 0.01 level of test would be rejected if the *p*-values were less than 0.005 under a Bonferroni correction. It appeared that both variables showed a significant year effect for all items. However, there was a comparatively weaker effect of design for five or six items.

Even though most of the F statistics were statistically significant, the magnitude of effect should also be taken into account. The eta-squares (η^2) were computed to measure the size of overall effect for design and years of administration. The values had a wide range from 0.03 to 0.79 for design effect and the maximum is 0.74 for year effect. However, in general the patterns between simulation data and real data were similar. There were no substantial differences among item parameter estimates.

Table 4.13 Results of MANOVA F-statistics, and η^2 for Model IV

	Overall					
Item	Design			Year		
	Λ	$F_{4,2391}$	η^2	Λ	$F_{8,4782}$	η^2
L01	0.87	85.4	0.13	0.60	173.4	0.22
L02	0.80	146.7	0.20	0.26	578.3	0.49
L03	0.67	295.1	0.33	0.16	874.3	0.59
L04	0.55	490.3	0.45	0.89	36.2	0.06
L05	0.92	51.9	0.08	0.39	363.5	0.38
L06	0.97	16.4	0.03	0.77	82.7	0.12
L07	0.44	762.5	0.56	0.35	417.5	0.41
L08	0.93	45.4	0.07	0.41	333.3	0.36
L09	0.89	75.6	0.11	0.40	351.5	0.37
L10	0.21	2299.4	0.79	0.35	405.6	0.40
G11	0.88	85.3	0.12	0.28	536.9	0.47
G12	0.83	126.4	0.17	0.68	128.8	0.18
G13	0.76	184.8	0.24	0.51	243.0	0.29
G14	0.88	83.1	0.12	0.62	159.4	0.21
G15	0.80	146.7	0.20	0.70	118.7	0.17
G16	0.29	1434.4	0.71	0.21	706.7	0.54
G17	0.69	267.0	0.31	0.42	320.2	0.35
G18	0.28	1533.8	0.72	0.13	1029.7	0.63
G19	0.79	155.7	0.21	0.30	492.1	0.45
G20	0.93	44.3	0.07	0.64	150.0	0.20
V21	0.35	1121.8	0.65	0.59	180.2	0.23
V22	0.52	546.7	0.48	0.75	90.7	0.13
V23	0.44	762.8	0.56	0.57	192.8	0.24
V24	0.75	203.6	0.25	0.47	277.7	0.32
V25	0.93	45.1	0.07	0.62	160.1	0.21
V26	0.81	140.2	0.19	0.46	283.2	0.32
V27	0.71	246.8	0.29	0.43	309.3	0.34
V28	0.58	436.8	0.42	0.11	1200.7	0.67
V29	0.64	339.1	0.36	0.07	1678.2	0.74
V30	0.87	93.0	0.13	0.81	65.6	0.10

Note: * These F statistics were statistically not significant at p value of 0.01 and univariate tests for *a* and *b* used Bonferroni correction at $p \geq 0.005$.

Table 4.14 Results of MANOVA test criteria for Model IV

Item	a_1		a_2		a_3		d	
	Design	Year	Design	Year	Design	Year	Design	Year
	$F_{1,2394}$	$F_{2,2394}$	$F_{1,2394}$	$F_{2,2394}$	$F_{1,2394}$	$F_{2,2394}$	$F_{1,2394}$	$F_{2,2394}$
L01	99.8	13.6	6.3*	8.6	4.9*	1.4*	40.2	668.6
L02	112.4	3.2*	0.8*	134.2	2.6*	2.6*	243.2	2569.5
L03	528.5	138.0	3.2*	38.1	26.6	62.6	1051.5	4305.5
L04	2.2*	30.4	54.4	18.8	1617.9	54.1	40.0	27.2
L05	6.8*	99.2	23.9	67.3	179.9	215.6	8.3	693.3
L06	11.7	230.6	0.1*	12.1	14.5	18.4	30.5	88.3
L07	26.1	102.6	0.1*	25.7	2537.8	301.6	306.7	1782.2
L08	13.2	342.6	13.1	142.6	67.5	55.1	34.6	490.8
L09	12.6	75.8	7.5*	1.5*	13.0	8.4	276.1	1637.5
L10	57.7	255.8	7222.1	47.4	124.1	41.2	56.0	1386.3
G11	37.2	3.9*	178.6	248.2	94.0	32.9	38.7	2112.1
G12	0.3*	26.4	152.1	190.8	359.7	22.7	70.4	385.3
G13	6.3*	61.2	385.1	239.9	394.7	28.6	27.7	555.2
G14	47.7	294.3	14.5	6.2	22.3	222.2	199.3	7.5
G15	126.5	17.2	288.8	18.2	0.1*	120.7	6.0*	369.3
G16	220.7	54.8	3153.6	859.7	4765.6	632.6	1077.9	3200.0
G17	473.9	83.2	439.8	476.1	0.1*	8.0	183.7	1312.3
G18	207.8	42.4	1409.1	34.4	2224.9	93.6	1527.7	6111.0
G19	53.0	171.7	127.8	71.2	349.2	150.0	143.7	1799.2
G20	11.3	26.7	1.5*	73.5	11.1	281.3	156.0	284.7
V21	8.3	347.3	2041.0	69.3	2228.7	11.0	1.9*	313.6
V22	272.3	100.6	1095.0	184.8	682.2	17.6	24.4	61.7
V23	490.7	39.2	1217.9	9.2	738.7	10.4	41.9	169.3
V24	53.5	3.3*	491.4	32.7	324.1	203.4	40.3	919.9
V25	13.5	20.9	98.3	62.2	53.3	44.9	7.5*	251.1
V26	4.1*	24.8	0.1*	26.9	113.8	43.4	379.1	1150.7
V27	25.8	96.4	291.0	700.4	763.7	131.7	93.0	419.8
V28	96.2	25.4	144.2	22.6	196.9	29.3	1409.5	9333.2
V29	4.1*	281.0	323.3	111.8	1228.0	158.2	9.9	11475.2
V30	13.3	39.2	17.9	136.4	149.0	68.0	229.4	42.4

Note: * These F statistics were statistically not significant at p value of 0.01 and univariate tests for *a* and *b* used Bonferroni correction at $p \geq 0.005$.

In summary, the mean distributions of both item difficulty and item discrimination were indistinguishable over time or by design. There was no substantive variation in item parameter estimates with 400 replications. The unidimensional models were robust in terms of recovering item parameters. No lack of invariance has been identified. Even the unidimensionality assumption was violated by using multidimensional test data; there was no substantial variation in item parameter estimates. However, the estimates of multidimensional models appeared to be more similar across replications than those of unidimensional models. In addition, item difficulties were relatively stable compared to the item discrimination values.

4.4 Comparison of Four Models Assuming Different Factor Structures

4.4.1 Item Parameter Recovery

Bias was evaluated separately for each item parameter for each model between simulation samples and samples selected from real data. Figures 4.15, 4.16, 4.17, 4.18 and 4.19 showed the results for Models I, II, III, and IV. Average bias for the 400 simulation samples generated using the same sets of item parameters provided the basis against which the bias estimation were compared from 400 samples of real data. As expected, the bias estimates for simulation data represented by dotted lines were close to zero lines. However, the average bias estimates from real data were widely spread and considerably large for a few items.

The plots on the upper row showed the average bias for *discrimination* parameters. The results of Model I was more stable than Models II and III. The average bias estimates that were consistently large across all models were for item 17, item 21, and item 22. Especially for Models II and III, two more items (Item 3 and item 8) also had considerably large bias values.

The plots on the bottom row showed the average bias for *difficulty* parameters. The patterns were different from those of the *a* parameters. Several items exhibited large bias in Model I but the values tended to decrease for Models II and III. However, a few items still had great variations over years, such as item 18, item 28, and item 29.

Compared to three unidimensional models, the multidimensional Model IV in Figures 4.18 and 4.19 had smaller bias for all item parameter estimates, which also showed less variation over years.

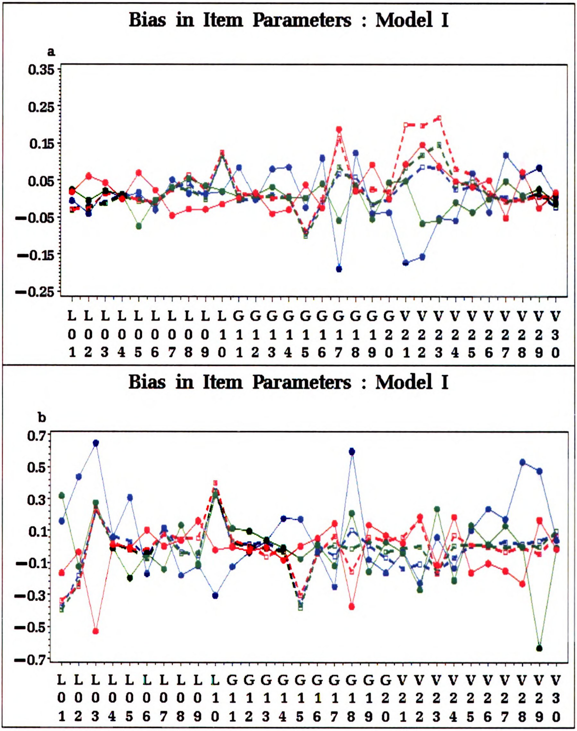


Figure 4.15 Bias for item parameter estimates for Model I[11]
(Images in this dissertation are presented in color)

[11] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

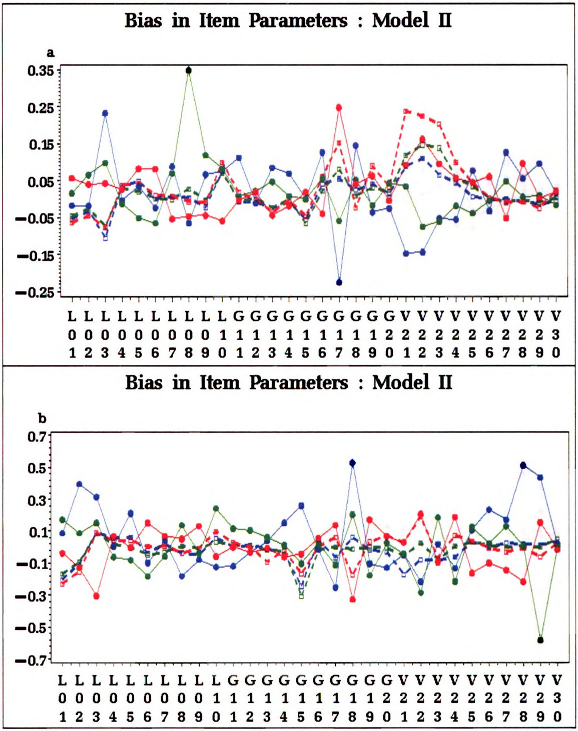


Figure 4.16 Bias for item parameter estimates for Model II[12]
(Images in this dissertation are presented in color)

[12] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

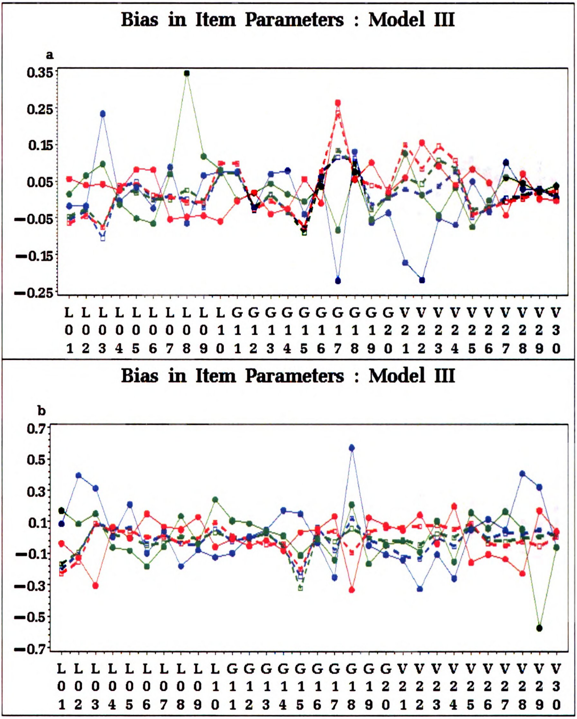


Figure 4.17 Bias for item parameter estimates for Model III[13]
(Images in this dissertation are presented in color)

[13] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

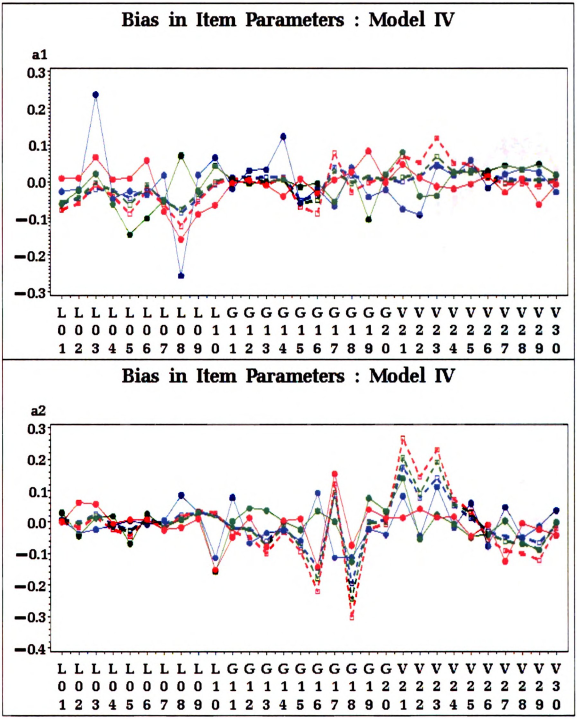


Figure 4.18 Bias for item parameter estimates for a_1 and a_2 for Model IV[14] (Images in this dissertation are presented in color)

[14] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

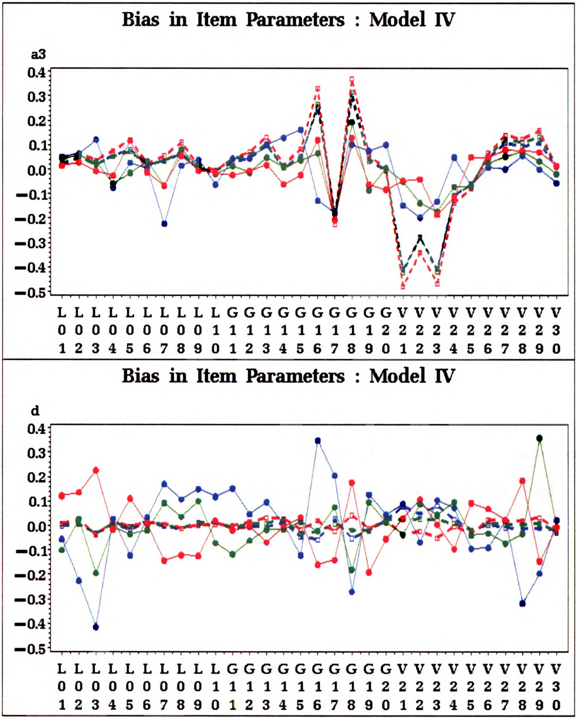


Figure 4.19 Bias for item parameter estimates for a_3 and d for Model IV[15]
(Images in this dissertation are presented in color)

[15] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

Root mean squared error (RMSE) was also evaluated separately for each item parameter for each model between simulation samples and samples selected from real data. RMSE was more meaningful compared to bias because it does not allow cancellation between the positive and the negative values. Figures 4.20, 4.21, 4.22, 4.23, and 4.24 exhibited the results for Models I, II, III, and IV. Average RMSE for the 400 simulation samples generated using the same sets of item parameters provided the basis against which the estimation from 400 samples of real data were compared. As expected, the bias estimates for simulated data represented by dotted lines were close to zero lines. However, the average RMSE estimates from real data had a wider range and were large for a few items.

The plots on the upper row show the average RMSE for *discrimination* parameters. The results of Model I were comparatively stable compared to Models II and III. The average RMSE estimates that were consistently large across all models were item 17, item 21, and item 22. Especially for Models II and III, two more items (Item 3 and item 8) also had considerably large bias values.

The plots on the bottom row showed the average bias for *difficulty* parameters. The patterns were different from those of the *a*-parameters. Several items exhibited large RMSE in Model I but the values tended to decrease for Models II and III. However, a few items still had large variations over years, such as item 18, item 28, and item 29.

Compared to three unidimensional models, the multidimensional Model IV in Figure 4.23 and 4.24 had smaller RMSE for all item parameter estimates, which also showed less variation over years.

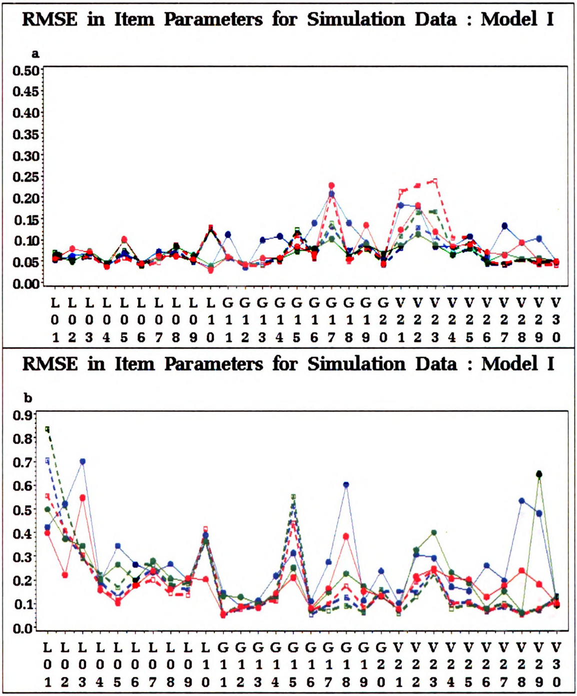


Figure 4.20 Root Mean Squared Difference (RMSE) of item parameter estimates for Model I[16]
(Images in this dissertation are presented in color)

[16] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

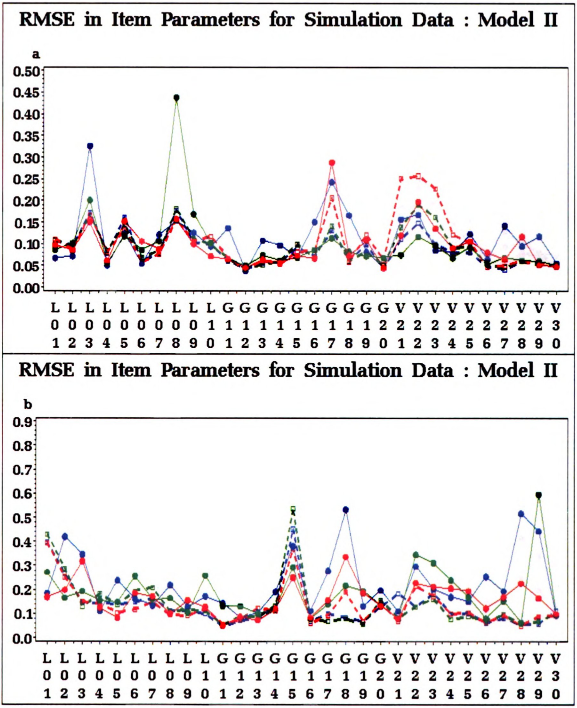
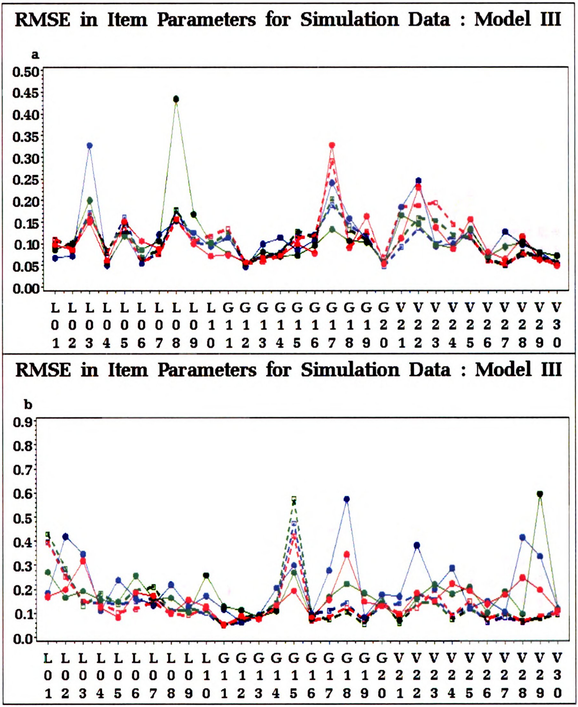


Figure 4.21 Root Mean Squared Difference (RMSE) of item parameter estimates for Model II[17]
(Images in this dissertation are presented in color)

[17] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.



RMSE in Item Parameters for Simulation Data : Model III

RMSE in Item Parameters for Simulation Data : Model III

Figure 4.22 Root Mean Squared Difference (RMSE) of item parameter estimates for Model III[18]
(Images in this dissertation are presented in color)

[18] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

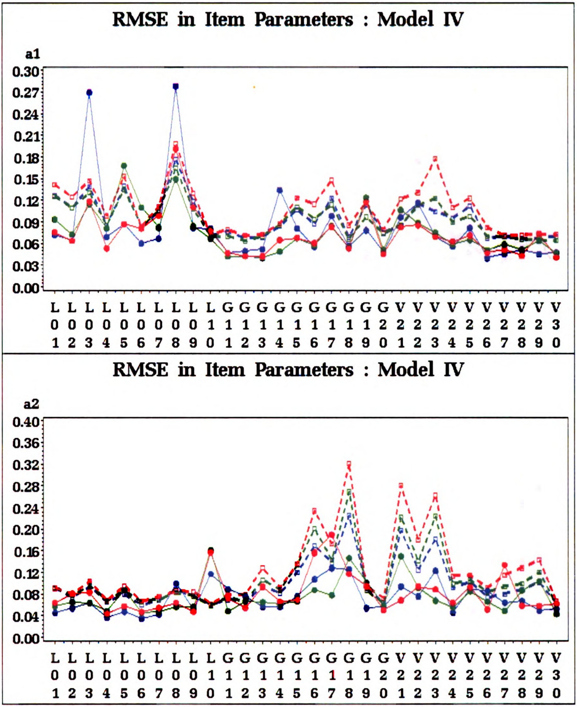


Figure 4.23 Root Mean Squared Difference (RMSE) of item parameter estimates for a_1 and a_2 for Model IV[19]
(Images in this dissertation are presented in color)

[19] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

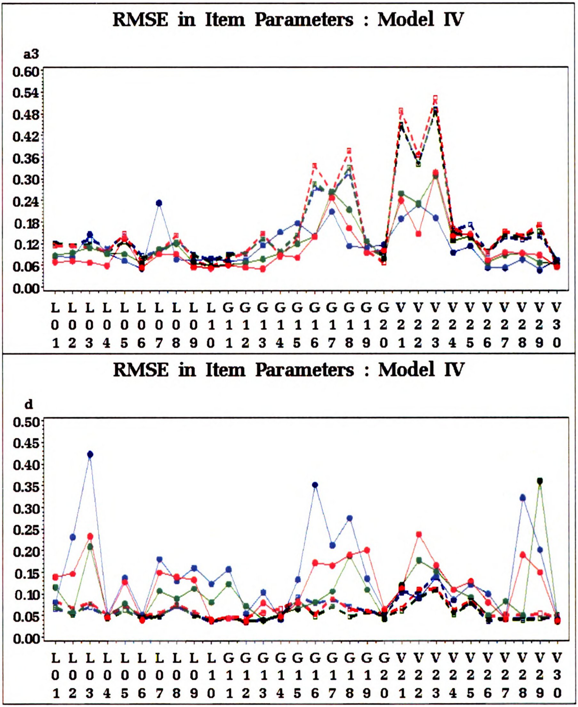


Figure 4.24 Root Mean Squared Difference (RMSE) of item parameter estimates for a_3 and d for Model IV[20]
(Images in this dissertation are presented in color)

[20] The dotted lines represent the parameter estimates based on the simulated sample and the solid lines represent the parameter estimates based on samples from real data. The lines and markers in blue represent item estimates from year 1. The lines and markers in green represent item estimates from year 3. The lines and markers in red represent item estimates from year 6.

4.4.2 Item Parameter Drift

Previous analyses revealed that item parameter estimates exhibited variation over time to some extent. However, further investigation was required to determine how the difference in item parameter estimates affected the true scores and resulted in significantly different item performance over time. The NCDIF index proposed by Raju, van der Linden, & Fleer (1995) was defined as a weighted measure of the squared difference between the item response functions.

Item parameter estimates were calibrated separately for simulation samples and real data samples. They were both linked to the same scale as the true parameters that were used for generating item responses. Finally, the NCDIF indexes were computed with the item parameter estimates for both groups set in models.

In this study, the null distribution of NCDIF was generated by sorting the 400 NCDIF values calculated using the simulation data. As assumed, the simulation data represented no drift situation except for measurement error. A cut-off value was then determined by obtaining the (1- α) percentile with the type I error rate of α. Given the choice of α values of 0.05 or 0.01, the 95% and 99% confidence interval for the distribution of NCDIF index were created.

The count of NCDIF index values that were greater than the cut-off values set were summarized in Tables 4.15, 4.16, and 4.17. Each table represented the results from one year of administration. Out of the 400 replications, the numbers showed the deviation of the distribution of NCDIF for real data from the distribution based on simulation data.

The null hypothesis was that the mean of NCDIF distribution from simulation samples was equivalent to that from real data samples. The larger the number was, the more frequent the NCDIF values in real data sample were rejected as being the same distribution as the null. Each model had two columns for two rejection values.

For the first year, the largest number of rejections was 400 (100%) times for item 2, item 18, item 28 and item 29 under Model I. The frequency tended to decrease as the number of factors increased in the models. In general, the overall frequency of rejection had a pattern of gradual decline. Especially for Model IV underlying MIRT models had only a few cases rejected. The α=0.01 case observed only one sample NCDIF being rejected.

For the third year, the largest number of rejections was 400 (100%) times for item 29 under Model I. The frequency tended to decrease as the number of factors increased in the models. In general, the overall frequency of rejection had a pattern of gradual decline. Especially for Model IV underlying MIRT models had only a few cases rejected. The α=0.01 level had only one case rejected for most of the items.

For the sixth year, the largest number of rejections was 398 (99.5%) times for item 28 under Model I, followed by item 18 with 389. The frequency tended to decrease as the number of factors increased in the models. In general, the overall frequency of rejection has a pattern of gradual decline. Especially for Model IV underlying MIRT models had only a few cases rejected. The α=0.05 level observed five items rejected a few times but no item had been rejected for the NCDFI index at a significance level of 0.01.

Table 4.15 Number of replications with NCDIF above the critical-value for Year 1

	Model I		Model II		Model III		Model IV	
Item	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01
L01	127	48	91	56	91	55	1	0
L02	400	400	399	394	399	394	1	0
L03	362	339	388	356	388	356	101	0
L04	44	15	14	7	14	7	0	0
L05	307	265	218	93	218	93	0	0
L06	26	10	45	19	45	19	3	0
L07	87	19	149	56	150	57	0	0
L08	123	62	158	64	158	64	12	0
L09	146	40	142	58	154	57	3	0
L10	16	5	191	100	191	102	50	0
G11	292	210	321	240	102	21	0	0
G12	13	3	12	2	11	2	0	0
G13	101	52	85	58	42	10	0	0
G14	133	43	109	40	114	28	1	0
G15	144	85	223	176	230	140	0	0
G16	260	186	208	93	17	7	0	0
G17	263	166	291	183	179	56	0	0
G18	400	400	400	400	400	399	0	0
G19	96	54	96	24	35	2	0	0
G20	32	17	18	12	19	5	123	13
V21	194	143	96	44	251	169	0	0
V22	42	21	30	14	128	67	0	0
V23	63	14	72	26	207	103	0	0
V24	79	25	65	20	198	57	0	0
V25	170	114	200	142	61	24	0	0
V26	349	315	350	328	155	99	0	0
V27	254	182	281	197	150	84	0	0
V28	400	400	400	400	400	399	0	0
V29	400	400	400	400	389	368	0	0
V30	18	4	33	11	58	23	3	0

Table 4.16 Number of replications with NCDIF above the critical-value for Year 3

	Model I		Model II		Model III		Model IV	
Item	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01
L01	135	45	142	108	142	108	10	0
L02	15	8	41	21	41	21	2	1
L03	78	26	154	84	153	87	13	0
L04	24	6	12	6	12	6	1	1
L05	69	21	57	8	56	8	1	1
L06	10	5	87	30	87	30	9	0
L07	261	159	257	132	237	133	1	1
L08	45	18	232	152	228	150	1	1
L09	208	132	177	92	177	91	29	0
L10	0	0	97	54	97	54	1	1
G11	166	98	165	100	58	20	1	0
G12	79	25	91	44	74	22	1	1
G13	46	36	75	43	45	26	1	1
G14	21	8	34	7	20	7	2	0
G15	10	5	11	5	6	3	1	1
G16	48	15	43	24	23	3	1	1
G17	69	24	74	33	41	11	1	0
G18	280	217	336	276	175	52	1	1
G19	227	139	247	147	246	166	6	0
G20	50	16	49	16	24	12	3	0
V21	9	2	1	0	75	31	1	0
V22	141	84	114	70	56	18	1	0
V23	0	0	0	0	1	0	1	0
V24	242	98	210	89	88	35	1	0
V25	22	8	31	16	76	13	1	1
V26	37	13	39	12	97	35	1	0
V27	120	48	118	50	197	132	1	1
V28	38	7	35	17	91	26	1	1
V29	400	400	400	400	400	400	1	1
V30	13	2	20	6	34	12	2	0

Table 4.17 Number of replications with NCDIF above the critical-value for Year 6

	Model I		Model II		Model III		Model IV	
Item	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01
L01	87	44	72	21	72	21	0	0
L02	184	80	172	97	172	96	0	0
L03	236	171	346	260	347	253	1	0
L04	15	6	18	7	18	7	0	0
L05	87	21	46	6	45	6	0	0
L06	36	21	119	96	119	96	7	0
L07	243	154	293	180	293	180	0	0
L08	42	10	92	19	82	18	1	0
L09	235	141	240	153	260	153	21	0
L10	0	0	13	4	13	4	7	0
G11	13	1	15	5	3	1	0	0
G12	25	11	33	15	45	20	0	0
G13	24	2	10	1	30	6	0	0
G14	25	5	17	7	37	15	0	0
G15	10	1	14	4	11	3	0	0
G16	66	30	85	38	46	5	0	0
G17	74	23	112	51	66	25	0	0
G18	389	369	391	361	398	391	0	0
G19	192	122	145	92	137	44	0	0
G20	39	8	18	3	19	3	9	0
V21	0	0	0	0	3	0	0	0
V22	27	9	22	7	76	20	0	0
V23	0	0	0	0	1	0	0	0
V24	118	37	83	27	78	35	0	0
V25	43	10	77	31	108	22	0	0
V26	158	108	199	128	143	55	0	0
V27	109	55	91	51	75	33	0	0
V28	398	389	398	394	373	277	0	0
V29	284	209	225	99	212	106	0	0
V30	7	2	7	2	15	5	0	0

CHAPTER 5

CONCLUSION, IMPLICATIONS, AND LIMITATIONS

5.1 Conclusions

This study empirically examined the effect of multidimensionality upon the invariance property of item parameter estimates of IRT models. The data from three years of administration of ECPE were used to investigate the potential drift of both item difficulty and item discrimination estimates for the same set of items. Only the common items were examined across three administrations within six years for a total of 72,277 examinees. Four models with varying dimensions were used to calibrate and link the test data that were sensitive to multiple dimensions. Samples selected from real data were compared to the simulated data generated under a multidimensional model.

Multiple dimensions were identifiable for the 30 common items used in ECPE even when traditional methods using eigenvalue analysis identified a single dimension. The results of residual analyses and item vector plots suggested that three dimensions provided optimal solutions. The measurement invariance test using confirmatory factor analysis also supported that three-factor model had the best fit to the data. The results also confirmed that the pattern of factor structures remain stable over years. The metric invariance hypotheses were supported for equal factor loadings but not for equal thresholds. The dimensionality might be underestimated by conventional techniques because they relied on a dominant dimension and shared variance. The multiple constructs were highly correlated but they measured different composites of English proficiency. A weak assumption of measurement invariance was also supported across years and the model fit improved when the number of dimensions increased.

Preliminary analysis comparing mean plots showed that the item parameter estimates for simulated samples remained stable as expected. Especially for the item difficulty estimates, the means were equivalent over time. There were a few items exhibiting small deviations for item discrimination values, which were attributed to measurement error. A consistent stability was observed for different models. The patterns were similar for both bias and RMSE plots. It should be noticed that the scale range was different between *a* parameters and *b* parameters. Given that item difficult values in IRT had a wider range than the item discrimination parameters, the relative bias and RMSE were lower for estimates of *b* parameters compared to *a* parameters.

Compared to simulation samples without IPD, the results of real data samples revealed a pattern of variation across administrations. However, the degree of variation in the IRT item parameter estimates gradually decreased as the dimensions of the model increased. There was also a decline in the number of items with dissimilar parameter estimates. The multidimensional model had relatively less variance for all item parameter estimates over time.

In general, the item difficulty indices exhibited a very high degree of invariance across samples, even for calibrations using the one-dimensional model. No obvious negative effect of multidimensionality on the invariance property was observed. The models with lower dimensions showed a tendency to have slightly less invariance estimates than the models with higher dimensions. The estimation of item difficulty parameters was robust for both unidimensional and multidimensional models. They also remained stable across administrations.

The item discrimination parameter estimates, however, had greater variation than the item difficulty values. The degree of invariance of item discrimination parameter estimates also increased steadily as the dimensions of models increased, implying that the IRT discrimination parameter estimates did not maintain a high degree of invariance for items sensitive to multiple dimensions. In addition, the number of items with dissimilar discrimination estimates decreased over time with an increase of dimensions.

Using the NCDIF index for statistical significance of invariant parameter estimates, the differences in true scores for both groups were compared. It showed that the differences were not due to random noise but led to different item characteristic curves. The count of NCDIF values greater than the cut-off values provided guidance for what degree of parameter variation was within acceptable limits.

As a result, the analyses of real test data presented as examples in this paper showed that there was evidence of the effect of multidimensionality on parameter invariance. Multidimensional models generally exhibited less variation than unidimensional models and even for models assuming three dimensions based on sections. The results showed that the choice of models for calibration and linking tended to have a large effect on the resulting IPD detection. The increase in the amount or magnitude of IPD among the linking items might be due to the inadequate considerations of dimensionality. For items that were sensitive to multiple dimensions, models with higher dimensions produced similar indices across forms and were consistently the best among the models. The observed IPD using unidmenional models might indicate that inadequate dimensionality was addressed.

5.2 Implications

The findings of this study have important implications for the ESL/EFL tests but may be insightful for study of other content areas, such as mathematics with more complex factor structure. The assumption of unidimensionality is critical for any IRT analysis. Traditionally, it is typical to claim that there exists only one "dominant" factor that influences the test performance based on the eigenvalues or scree plots. However, this assumption cannot be completely met by any test data. The conventional exploratory factor analysis might be misleading especially for the test data with highly correlated factors. The findings of other studies are actually based upon a combination of measures that are the aggregate of multiple measures. The results are likely to mask what should be differential results related to invariance of item parameters. Researchers and practitioners should be cautious about assuming unidimensionality and the property of parameter invariance might be misleading when the assumption violated.

In addition, assessments apply valid and reliable techniques to make a fair evaluation. IPD poses a threat on the validity of scores by introducing trait-irrelevant differences in anchor items over time. The cut scores are determined by comparing the performance of linking items from one year to more previous years' tests. Failing to identify IPD can disadvantage individual test-takers and jeopardize test interpretations. However, misspecification of IPD due to dimensionality may also provide flawed information when generalized to other conditions. Thus, a better understanding of the dimensionality of the real data analyzed may lead to valid conclusions drawn from the interchangeable use of alternate forms, which would be valuable and helpful for practitioners in enhancing the quality of large-scale assessments.

5.3 Limitations

The results must be considered in light of study limitations. A variety of sources of error were expected for item calibration. These included estimation error due to statistical methods, sampling error because only selected samples were examined from the target population estimation, and measurement error consisting of random error and systematic error. The linking items were only a small part of the whole test. Analysis using only linking items was likely to increase estimation error with shorter tests or lead to additional sampling error due to different sets of common items selected.

Though measurement invariance was weakly achieved across years, it would be interesting to check whether the structure remained the same across countries where the test was administered. However, the test has been administered in about 120 testing centers across over 20 countries. The number of examinees for some countries was quite small and the model would not be identified. Another option was to check the measurement invariance by group based on the first language of examinees. Since the Greek account for about 95% to 98% of the target population of examinees for this test, it would also be very hard to examine those non-Greek groups. However, it might be interesting to combine those subgroups with small sample size that were categorized by either country or language background and test whether the dimensional structure differed for further exploration in the future.

This study illustrated to some extent the robustness of item response theory when the assumption of unidimensionality was violated. Though evidence of drift of item parameters was found, other factors might confound the sources of IPD. There was possibility that other variables might have confounding impact upon the incidence of

IPD, such as the location of linking items, the number of common items, and the complexity of the factor models. These topics could not be fully covered in this study but they might be areas that direct more research. In addition, the methodology applied in this study was similar to the statistical technique of bootstrap. A comparison of both procedures should be conducted in the future to check whether the same results could be replicated. For instance, some items had large magnitudes of IPD when they were administered in two different locations in the test, especially the end-of-test items used for linking (Wollack, et al, 2006; Oshima, 1994). Some degree of parameter variation might be reasonable indicating inherent changes in characteristics. However, the question was at what point it might lead to measurement nonequivalence and cause error in linking and equating. The crucial question was at what point parameter variation becomes critical and leads to biased results. Future research may attempt to specify models more precisely using a wider range of variables, with better measurement, that may result in stronger prediction of the sources of drift in item parameters.

APPENDIX A:

Code for Generating Scree Simulation Plots in MATLAB

```
function [G]=PA(D, P, N, I, K)
% PA: Parallel analysis generating random numbers based on a binomial
% distribution with the proboability equals to the probability of correct responses
% for each item each time
% D is the item response matrix
% P is the vector of probability of correct response for all three years
% P1 is the vector of probability of correct response for Year 1
% P2 is the vector of probability of correct response for Year 3
% P3 is the vector of probability of correct response for Year 6
% N is the simulated sample size
% I is the total number of items
% K is the number of replications

% Principal component analysis for real data samples
I=30;
K=400;
F=corrcoef(D);
[pc, latent] = pcacov(F);
T=latent;

% Principal compent analysis for simulation data
R= zeros(N,I);
E= zeros(I,K);
for k=1:K
        for i=1:I
                R(:,i)= binornd(1,P(i),N,1);
        end
        C=corrcoef(R);
        [pc, latent] = pcacov(C);
        E(:, k)=latent;
end
G= [T, E];

% Generate scree simulation plot for data from all three years.
M0=dlmread('All.dat');
D0=M0(:,1:30);
N0=size(D0,1);
P0=[0.9,0.847,0.578,0.751,0.833,0.726,0.815,0.841,0.767,0.583,0.525,0.587,0.684,0.778,
0.922,0.7,0.736,0.672,0.586,0.741,0.523,0.862,0.387,0.675,0.461,0.625,0.619,0.621,0.63
5,0.302;];
```

```
P0=P0';
G0=PA(D0,P0,N0);
plot(G0,'-ko','LineWidth',2,
        'MarkerEdgeColor','k',
        'MarkerFaceColor','w',
        'MarkerSize',4)
ylabel('Eigenvalue')
xlabel('Factors')
title ('Scree Simulation Plot for Three Years ')

% Generate scree simulation plot for data from Year 1.
M1=dlmread('Y1.dat');
D1=M1(:,1:30);
N1=size(D1,1);
P1=[0.892,0.803,0.528,0.755,0.812,0.739,0.834,0.868,0.796,0.624,0.576,0.607,0.711,0.7
82,0.911,0.745,0.777,0.595,0.621,0.761,0.559,0.875,0.415,0.705,0.453,0.602,0.619,0.533
,0.573,0.313;];
P1=P1';
G1=PA(D1,P1,N1);
plot(G1,'-ko','LineWidth',2,
        'MarkerEdgeColor','k',
        'MarkerFaceColor','w',
        'MarkerSize',4)
ylabel('Eigenvalue')
xlabel('Factors')
title ('Scree Simulation Plot for Year 1 ')

% Generate scree simulation plot for data from Year 3
M2=dlmread('Y3.dat');
D2=M2(:,1:30);
N2=size(D2,1);
P2=[0.895,0.859,0.557,0.765,0.851,0.737,0.848,0.848,0.801,0.567,0.522,0.588,0.703,0.7
91,0.934,0.727,0.764,0.659,0.633,0.761,0.543,0.889,0.4,0.717,0.467,0.639,0.62,0.643,0.7
67,0.323;];
P2=P2';
G2=PA(D2,P2,N2);
plot(G2,'-ko','LineWidth',2,
        'MarkerEdgeColor','k',
        'MarkerFaceColor','w',
        'MarkerSize',4)
ylabel('Eigenvalue')
xlabel('Factors')
title ('Scree Simulation Plot for Year 3 ')

% Generate scree simulation plot for data from Year 6
M3=dlmread('Y6.dat');
```

```
D3=M3(:,1:30);
N3=size(D3,1);
P3=[0.907,0.862,0.618,0.739,0.832,0.712,0.783,0.823,0.729,0.572,0.501,0.576,0.657,0.7
68,0.92,0.658,0.695,0.72,0.536,0.717,0.491,0.837,0.363,0.632,0.461,0.627,0.617,0.651,0.
578,0.281;];
P3=P3';
G3=PA(D3,P3,N3);
plot(G3,'-ko','LineWidth',2,
        'MarkerEdgeColor','k',
        'MarkerFaceColor','w',
        'MarkerSize',4)
ylabel('Eigenvalue')
xlabel('Factors')
title ('Scree Simulation Plot for Year 6 ')
```

APPENDIX B:

Code for Computing DTF/NDIF Index Values

```
function lkout(R,Y,I,J,t)
% LKOUT applied Raju's DTF/NDIF method to examine item parameter drift
% using oblique procrust rotation methods for simulation data.
% Ad is the TRUE item parameter used for 3 dimensional 3PL model.
% However, c parameters were calibrated by BIOLG-MG and fixed for both
% unidimensional and multidimensional calibration.
% TRUE parameters were based on calibration of the population of
% with three years combined.
% Y is the year of administration.
% R is the replication of the dataset.
% I is the number of items.
% J is the number of subjects. 2000 cases were used for both simulation and real data for
% each calibration.
% t is the target item parameter scale matrix
% OUT returns the DTF, CDIF, NDIF values and chi-square using Raju's DIF
% method to test the significance of the difference between the TRUE parameter and
% the rotated parameter estimates

% Read item parameter estimates from TESTFACT parameter file.
for y=1:Y
  for r=1:R
  fid=fopen(['y,num2str(y),'_r',num2str(r),'.par'],'r');
  K=textscan(fid,'%d %s %9.5f %9.5f %9.5f %9.5f %9.5f','headerlines',1);
  fclose(fid);
  f.c=[K{3}];
  f.d=[K{4}];
  f.A=[K{5},K{6},K{7}];
  f.A1=[K{5}];
  f.A2=[K{6}];
  f.A3=[K{7}];

% Read theta matrix from TESTFACT output file.
fid=fopen(['y',num2str(y),'_r',num2str(r),'.fsc'],'r');
    for j=1:J
      fgetl(fid);
      tline=fgetl(fid);
      num=str2num(tline);
      theta(j,:)=num(1:3);
      fgetl(fid);
    end
  fclose(fid);

% Store the scaled paramter estimates and DFIT statistics.
X = out(f,t,theta,I,J);
fid=fopen(['y',num2str(y),'_r',num2str(r),'.OUT'],'W');
    for i=1:I
      for m=1:6
        fprintf(fid,'%9.5f %9.5f %9.5f %9.5f %9.5f %9.5f',X(i,m));
      end
      fprintf(fid,'\n');
    end
  fclose(fid);
  end
end

function [X]= out(f,t,theta,I,J)

%  Oblique procruste rotation
  T=inv(f.A'*f.A)*f.A'*t.A;                % Rotation matrix
  h.A=f.A*T;                               % A paramteres after rotation
  S= f.A*inv(f.A'*f.A)*f.A'*(t.d-f.d);     % Transformation  matrix
  h.d=f.d+S;                               % d paramter after transformation
  h.Ad=[h.d,h.A];
  h.A1=h.A(:,1);
  h.A2=h.A(:,2);
  h.A3=h.A(:,3);
  M=inv(t.A'*t.A)*t.A'*(f.d-t.d);        % Transformation matrix
  thetah=(inv(T)*theta'+M*ones(1,size(theta,1)))';

% Raju's Differential Item and Test Functioning
  t.p=zeros(J,I);
  f.p=zeros(J,I);
  O=ones(J,I);
  C=ones(J,1)*f.c';
  rZ=1.702*(thetah*t.A'+ones(J,1)*t.d');
  fZ=1.702*(thetah*h.A'+ones(J,1)*f.d');
    for i=1:I
    t.p(:,i)=C(:,i)+(O(:,i)-C(:,i)).*exp(rZ(:,i))./(1+exp(rZ(:,i)));
    f.p(:,i)=C(:,i)+(O(:,i)-C(:,i)).*exp(fZ(:,i))./(1+exp(fZ(:,i)));

% Compute true score for each examinee assuming it being member of either
% reference or focal group
    t.TCC=sum(t.p,2);
    f.TCC=sum(f.p,2);

% Compute difference between true scores for the whole test
```

```
    DD=f.TCC-t.TCC;
    mD=mean (DD);
% Compuate difference between true scores for each item
    dd=f.p-t.p;
    dsquare=dd.^2;
    md=mean(dd,1);
    NCDIF=mean(dsquare)';
    DF(:,i)=(1/(J-1)).*(DD-mD)'*(dd(:,i)-md(:,i))+mD*mean(dd(:,i));
    CDIF=(DF)';
    end
 d=h.d;
 A1=h.A1;
 A2=h.A2;
 A3=h.A3;
 ND=NCDIF;
 CD=CDIF;
 X=combine(A1',A2',A3',d',CD',ND');
 X=X';
```

REFERENCES

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

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issue and Practice, 22* (3), 37-51.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing.* Washington, D.C.: Author.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. *Applied Psychological Measurement in Education, 1,* 215-222.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading MA: Addison-Wesley.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275-285.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K.A. & Long, J.S. (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Chan, K.-Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Educational Measurement, 84,* 610-619.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practices, 10* (3), 37–45.

Cook, L. L., Eignor, D.R., Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*(1), 31–45.

Davey, T., & Parshall, C. G. (April 1995). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

De Boeck, P., Wilsom, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112* (1), 129-158.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17,* 265-300.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22,* 33-51.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*, 3-17.

Dorans, N. J. (2000). *Distinctions among classes of linkages* (Research Notes RN-11). New York: The College Board.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227-246.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281-306.

Drasgow, F., & Lissak, R.I. (1983) Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections* (Research Report 85-10). Princeton, NJ: Educational Testing Service.

Fan, X., & Ping, Y. (April, 1999). *Assessing the effect of model-data misfit on the invariance property of IRT parameter estimates.* Paper presented at Annual Meeting of the American Educational Research Association, Montreal, Canada.

Fraser, C. (1993). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory (Version 2). Armidale, New South Wales, Australia: The University of New England Center for Behavioral Studies.

Goldstein, H. A. R. V. (1983). Measuring changes in educational attainment over time: problems and possibilities. *Journal of Educational Measurement, 20,* 369-377.

Gorsuch, R. L. (1983). *Factor Analysis*. Lawrence Erlbaum Associates.

Hambleton, R. K. & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., & Rovinelli, R. J. (1986).Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Hingham, MA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., Zaal, N. J., & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & N. J. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 341-366). Boston: Kluwer Academic.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Publishing.

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

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

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, *2* (2), 141-154.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Eds.), *Educational Measurement* (4th ed., pp.187-220). Praeger.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood IL: Dow Jones-Irwin,

Jang E. E. & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*(1), 1-21.

Johnson, J. S., Li, X., Yamashiro, A. Y., & Yu, J. (2006a). *The ECPE annual report: 2001-2002* Ann Arbor, MI: ELI-UM.

Johnson, J. S., Li, X., Yamashiro, A. Y., & Yu, J. (2006b). *The ECPE annual report: 2004-2005* Ann Arbor, MI: ELI-UM.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions.* Thousand Oak: Sage Publications.

Kaskowitz, G. S., & Ayala, R. J. D. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25* (1), 39-52.

Kelkar, V., Wightman, L. F., & Luecht, R. M. (2000, April). *Evaluation of the IRT parameter invariance property for the MCAT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Kingston, N. M. & Dorans, N. J. (1984) Item location effects and their implications for IRT equating and adaptive Testing. *Applied Psychological Measurement,* 8 (2), 147-154

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

Kolen, M. J. & Brennan, R. L. (1995) *Test Equating: Methods and Practices.* New York: Springer-Verlag.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

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

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*(2). Available online: http://pareonline.net/getvn.asp?v=12&n=2

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6*, 463–472.

Lord, F.M., & Novick, M.R. (1968) *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Martineau, J. A. (2004). *The Effects of Construct Shift on Growth and Accountability Models.* Unpublished Dissertation, Michigan State University, East Lansing.

Martineau, J. A., Subedi, D. R., Ward, K. H., Li, T., Diao, Q., Drake, S., Kao, S., Li, X., Lu, Y., Pang, F., Song, T., & Zheng, Y. (October, 2006). *Non-linear unidimensional scale trajectories through multidimensional content spaces: a critical examination of the common psychometric claims of unidimensionality, linearity, and interval-level measurement.* Paper presented at the conference entitled Assessing and Modeling Cognitive Development in School: Intellectual Growth and Standard Setting, College Park, Maryland.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical Statistical Psychology, 34,* 100-117.

McDonald, R. P. (1999).*Test theory: A Unified Approach.* Mahwah, NJ: Lawrence Erlbaum.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99-114.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*(1), 23-40.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525-543.

Mislevy, R. J. (1982, March). *Five steps toward controlling item parameter drift.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Mulaik, S. A. (1972). *The Foundations of Factor Analysis.* New York: McGraw-Hill.

Muraki, E., & Bock, D. (2003). PARSCALE: IRT Scaling, Item Analysis, and Scoring or Rating Scale Data (Version 4.1). Chicago, IL: Scientific Software International.

Muthén, L. K., & Muthén, B. O. (2001). Mplus user's guide (Version 2). Los Angeles, CA: Muthén & Muthén.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41-68.

Oshima, T. C. (1994). The effects of speededness on parameter estimation in item response theory models. *Journal of Educational Measurement, 31,* 200-219.

Oshima, T.C. & Miller, M.D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. *Journal of Educational Measurement,* 27, 273-283.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34,* 253-272.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics, 4*, 321-334.

Recakse, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 3,* 207-230.

Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D. (1994). What is the "correct" dimensionality for a set of item response data? in D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 87-92). Ottawa, Ontario, Canada: University of Ottawa.

Reckase, M. D. (2006). Multidimensional Item Response Theory. *Handbook of Statistics: Psychometrics, 26,* 607-642.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25* (3), 193-203.

Reckase, M. D., & Martineau, J. A. (2004, October). *Growth as a multidimensional process.* Paper presented at the Annual Meeting of the Society for Multivariate Experimental Psychology, Naples, FL.

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

Roussos, L.A., Stout, W.F., & Marden, J. I. (1998). Using new Proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*(1), 1–30.

Rupp, A.A., & Zumbo, B.D., (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

Steinberg, L. & Thissen, D. (1995). Item response theory in personality research. In P. Shrout & S. Fiske (Eds.), *Personality research, methods & theory: A Festschrift honoring Donald W. Fiske* (pp. 161-181). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stenbeck, M. (1992). Review: Fundamentals of item response theory. *Contemporary Sociology, 21*(2), 289-290.

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

Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4,* 125-141.

Stone, C. A. & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams an empirical comparison of methods using the multistate bar examination. *Educational and Psychological Measurement, 66*(2), 193-214.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 55* , 293-325.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST manual.* Unpublished manuscript, University of Illinois, Urbana-Champaign.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). New York: Springer.

Stout, W. F., Nandakumar, R., Junker, B., Chang, H., & Steidinger, D. (1991). DIMTEST: A Fortran program for assessing dimensionality of binary item responses. *Applied Psychological Measurement, 16,* 236.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariancebased nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331-354.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement, 29,* 201-211.

Sykes, R. C., & Ito, K. (1993, April). *Item parameter drift in IRT-based licensure examinations.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Tam, H. P. & Li, Y. H. (1997, March). *Is the use of the difference likelihood ratio chi-square statistics for comparing nested IRT models justifiable?* Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & T. S. Long (Eds.), *Testing Structural Equation Models.* Newbury Park, CA: Sage.

Tate, R. L. (2002). Test dimensionality. In J. Tindal & T. M. Halayyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis.* Mahwah, NJ: Lawrence Erlbaum.

Tate, R. L. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27,* 159-203.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3* (1), 4-70.

van Abswoude, A.A.H., van der Ark, L.A., & Sijtsma, K., (2004) A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*(1), 3-24.

van der Linden, W.J. & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local item dependence on reliability? *Educational Measurement: Issues and Practice, 15,* 22-29.

Walker, C.M., Razia, A., & Thomas, S. (2006). Statistical versus substantive dimensionality: the effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement, 66*(5), 721-738.

Wang, M. (1987, April). *Estimation of ability parameters from response data to items trait are precalibrated with a unidimensional model.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wang, M. (1988, April). *Measurement bias in the application of a unidimensional model to multidimensional item-response data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Way, W. D., Way, Garey, & Golub-smith (1992). *An exploratory study of characteristics related to IRT item parameter invariance with the test of English as a Foreign Language.* (TOEFL Technical ETS-RR-92-43; ETS-TR-6). Princeton: Educational Testing Service.

Weiss, D.J. & Yoes, M.E. (1991). Item Response Theory. In R.H. Hambleton and J.N. Zaal (Eds.), *Advances in educational and psychological testing.* (pp. 69-95). Boston: Kluwer Academic Publishers.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77-87.

Wilson, D. T., Wood, R., Gibbons, R., Schilling, S. G., Muraki, E., & Bock, R. D. (2003). TESTFACT: Test scoring and full information item factor analysis (Version 4.0), Chicago, IL: Scientific Software International.

Witt, E.A., Stahl, J.A., Bergstrom, B.A., & Muckle, T. (2003) *Impact of item drift with non-normal distributions.* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). *The impact of compounding item parameter drift on ability estimation.* Paper presented at Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Yu, C. H., & Popp, S. E. O. (2005) Test equating by common items and common subjects: concepts and applications. *Practical Assessment Research & Evaluation,* 10 (4). Available online: http://pareonline/getvn.asp?v=10&n=4

Zhang, J., & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129-152.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows*. Chicago, IL: Scientific Software International.