



## **An Investigation of the Item Parameter Drift in the Examination for the Certificate of Proficiency in English (ECPE)**

**Xin Li**  
Michigan State University

**ABSTRACT** Common items are widely used to equate test scores of multiple forms or to link scores at different educational levels. Violation of assumptions of IRT models may challenge parameter invariance of these linking items for repeated use over time. Unidimensionality is one of the primary IRT assumptions. However, linking items are meant to be representative of the whole test and are likely to be sensitive to multiple dimensions that may exist. Parameter drift occurs if the statistical properties of items change over time (Goldstein, 1983). In this study, random samples were selected from the Examination for the Certificate of Proficiency in English (ECPE), of which 30 linking items were administered three times within six years. Simulated item responses for these common items were generated using parameter estimates concurrently calibrated for the entire population with three administrations combined. In comparison with the simulated data assuming no drift in item parameters, the property of parameter invariance for real data samples were examined for different latent trait distributions at different time points when the items were calibrated and linked using both unidimensional and multidimensional techniques. The results confirm that the potential effect of multidimensionality was found associated with the item parameter drift (IPD) for the same set of common items using four different data dimensional compositions. Also discussed are limitations for the constructed compositions of actual item responses and the robustness of IPD detection. Future areas of research are suggested.

In large-scale assessment, it has been a common practice to insert a set of items into operational tests for repeated use across years or over multiple administrations. These common items, referred to as anchor or linking items, are used to equate test scores of multiple test forms or to link tests at different educational levels. Test scores are converted to the same scale and become comparable for different groups of examinees. 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 anchor items change over time (e.g., item difficulty value and item discrimination power), which is referred to as item parameter drift (IPD) (Goldstein, 1983).

Drift is likely to occur when maintaining an item pool over time even though good-quality items are selected and secured carefully. Such effects may be expected because of frequent item exposure, increasing practice effect, or inappropriate test-wise training. Items may also perform differently across years due to changes in the construct or content. In 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. In particular, changes related to national, ethnic, and cultural issues may confound estimates of item parameters for English-as-a-second or foreign language (ESL/EFL) tests.

Although item statistics based on classical test theory (CTT) can measure the difficulty level and discrimination power of any item, they have been generally recognized as sample dependent. As an alternative, item response theory (IRT) models have the property of invariance for item and ability parameters given the model fit to the test data of interest (Hambleton & Swaminathan, 1985). In 1991, Hambleton, Swaminathan, and Rogers stated that “ability estimates obtained from different sets of items” and “item parameter estimates obtained in different groups of examinees will be the same (except for measurement errors)” (p. 8).

The features of invariant item parameters are primarily desirable in maintaining an 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. Score scale conversions are derived from the responses to those embedded common items, which assume the parameters that characterize linking items are independent of ability distribution of the examinees over multiple administrations. The unchanged item parameters make the observed difference in scaled scores attributable to the difference in abilities across groups or measurement of growth over time. Given the importance of invariant item parameters and unidimensionality, it is logical to expect that changes in item parameters may pose a threat for measuring the latent construct. Drift of anchor item parameters in particular may severely jeopardize a fair score conversation using linking items over multiple administrations, which may lead to false decisions in certification and licensure test.

In the 1980s researchers introduced the concept of IPD (Bock, Muraki, & Pfeiffenberger, 1988) to represent the changes in item parameters over time and found one potential source of IPD was curriculum. 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. An example was a fourth-grade science test item about the metric system, which 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 English system, which resulted in declining difficulty for items concerning the metric system but increasing difficulty for the English system items. 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 could be attributed to a shift in the physics curricula. Similar studies were also conducted in the field of applied psychology. Chan, Drasgow, & Sawin (1999) found an 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 compared to other item types. The authors called for attention to IPD studies and regular checks for possibility of IPD.

A common approach suggests that a linear relationship should be checked for item parameter estimates over time (Hambleton & Swaminathan, 1985; Lord, 1980). During the past decades, researchers have been concerned with finding ways to identify the potential threats of parameter invariance from item bias across subgroups, namely differential item functioning (DIF), though relatively few studies have examined changes in item parameters over repeated administrations. Similar statistical procedures can be used to assess both. As suggested by Angoff (1988), 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. Donoghue & Isham (1998) compared a number of DIF measures for detecting IPD. However, their simulated data only covered two time points with one year apart, which was also the situation in Wells, Subkoviak, and Serlin (2002) and Stone and Lane (1991). Two time points are used for applying methods to compare two subgroups in most DIF analyses, but this might not be sufficient to examine IPD, and may not be generalizable to multiple time points.

In reality it is typical to expect IPD over multiple testing occasions (Wollack, Sung, & Kang, 2006). Several studies of IPD have multiple time points over a span of more than four years (Bock et al., 1988). However, most of them considered drift in item difficulty only (Davey & Parshall, 1995; Sykes & Fitzpatrick, 1992; Sykes & Ito, 1993). A few studies identified changes in both item difficulty and item discrimination (DeMars, 2004; Chan et al., 1999). Even though several proposed the three-parameter logistic model, they were limited to two parameters for simplified interpretation. (DeMars, 2004; Donoghue & Isham, 1998).

A review of the literature reveals that a variety of IRT models is used for investigating IPD (DeMars, 2004; Kelkar, Wightman, & Luecht, 2000; Sykes & Ito, 1993). Unidimensionality is the underlying assumption behind IRT, that is, it is assumed that a single construct or trait is measured by a set of items. The assumption is easily violated due to the multidimensional nature of test items and test purposes in educational and psychological tests. Even though it is theoretically assumed that any application of IRT models requires unidimensional data, it is still unclear what effect multidimensionality has on the invariance property of IRT parameter estimates. Multidimensionality has been found to affect item parameter estimates, which consequently influenced item characteristic curves and true scores (Oshima, Raju, & Flowers, 1997). However, there is a lack of empirical research about the potential consequences of test data sensitive to multiple dimensions on the invariance property of IRT parameter estimates. The validity of IRT-based techniques might deteriorate to the degree that the data do not meet the assumption of the model (Oshima et al., 1997). As a result, a comparison of analyses using unidimensional and multidimensional IRT models is necessary for detection of IPD, and the impact of multidimensionality on IPD must be further explored.

The primary purpose of this study is to examine empirically the potential effect of dimensionality on parameter invariance of linking items in tests across multiple administrations. To be specific, the research question addressed is whether the parameter estimates of the anchor items differ over cycles and to what extent the invariance property of IRT item parameter estimates is threatened by the violation of unidimensional IRT models. Item parameter drift is investigated in the context of a large-scale certification test administered over years using both unidimensional and multidimensional techniques. A variety of test structures is compared by analyzing different combinations of sections from the real test data to identify the impact of multidimensionality on IPD detection.

## Method

### Data

The data are from the Examination for the Certificate of Proficiency in English (ECPE), which is a large-scale certification test of English as foreign or 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. Multiple-choice items are used for the listening and grammar/cloze/vocabulary/reading (GCVR) sections. All test forms follow the same clearly specified standardized procedures for each administration. The numbers of items administered for the listening and GCVR sections are 40 and 140. The numbers have changed to 50 for the listening section and 120 for GCVR sections, of which 10 grammar items and 10 vocabulary items are trial items and excluded from the final scores. In addition, 30 scored linking items are inserted into existing tests for equating to test scores of other forms. These items are from three sections: 10 in listening (L), 10 in grammar (G), and 10 in vocabulary (V). Cloze and reading items are not used as anchor items due to security issues.

In this study, data are analyzed using three administered forms that are labeled as Year 1, Year 3, and Year 6, respectively. The same set of common items was used for these three forms. Only responses for these 30 common items were included and there were no blank responses. Correct responses are scored as “1” and incorrect responses as “0”. The total number of examinees for all three administrations is 72,277.

### 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 that have eigenvalues greater than one (Kaiser, 1960), account for at least 10% of the total variance (Hatcher, 1994), and precedes a significant drop in a scree plot. Because all 30 common items are multiple-choice questions, the guessing parameter is included in the model to control for the probability of correct responses by the examinees with extreme low ability levels. However, guessing cannot be corrected for in common exploratory and confirmatory factor analysis models. With the assumption that these sections are highly correlated, exploratory analyses using an oblique rotation of loadings were conducted in both TESTFACT 4.0 (Wilson, Wood, Gibbons, Schilling, Muraki, & Bock, 2003) and NOHARM 2.1 (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 and NOHARM programs for calibration because NOHARM and TESTFACT cannot estimate guessing parameters.

The first step was to compute the eigenvalues based on the tetrachoric correlations using TESTFACT to identify important factors. In addition to the absolute values of eigenvalues, the relative change in eigenvalues for consecutive factors is proposed by Hattie (1985) to determine the number of dominant factors. The ratio of differences in consecutive factors, denoted as Factor Difference Ratio Index (FDRI) (Johnson, Li, Yamashiro, & Yu, 2006a), reflects the relative change in eigenvalues. Hattie (1985) suggests the ratio of the difference between the first and the second factor and the difference between the second and the third can be used to check the relative strength of the first factor. A rule of thumb proposed by Hattie (1985) is that a ratio greater than three is considered as a large difference in the 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 of residual statistics. Different statistics were computed by both TESTFACT and NOHARM but 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 were no greater than 0.05, indicating a close approximation between observed and expected values. Alternatively, the Root Mean Square of Residuals (RMSR), based on the difference between the observed item correlations and those implied by models, should be 0.05 or less for an acceptable factor solution (Muthen & Muthen, 2001). RMSRs are also available in NOHARM, but the residual statistic is based on the difference between the observed and model-based proportions of correct responses for each pair of items. As suggested by McDonald and Mok (1995), the criteria for this statistic is to be equal to or less than four times the reciprocal of the square root of the sample size to indicate fit to the data.

Finally, plots of the item vectors were used to decompose the 30 linking items in terms of the essential dimensionality. Proposed by Reckase and Ackerman (Ackerman, 1994; Reckase, 1985), item vector plots are scatterplots of item difficulties using an oblique factor analysis solution with three factors. The lengths of vectors are proportional to the multidimensional discrimination while the origin of the vectors indicate the three-dimensional item difficulties. A good indication of one essential dimension is when the item vectors fall on a straight line.

Previous studies have shown that the whole test has a dominant factor that is overall English skill. However, factor analyses also show a clear pattern of structure as items within the same test section tend to load high on the same factor except for a few cases (Johnson, Li, Yamashiro, & Yu, 2006a, 2006b). In this study, the data are 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 are tested and the effects of these different dimensionality structures on IPD are explored.

## **Design**

The item parameter estimates from concurrent calibration of the entire population, including all three years of administration, are taken as “true” item parameter values. The mean and variance-covariance matrices for abilities are estimated with the item parameters fixed. Based on these “true” item parameters and ability distributions of each year, simulated data are generated using the three-parameter three-dimensional logistic model. A sample of 2000 examinees’ responses is simulated for each administration year, and this process is replicated 400 times. The distributions of these simulated data involve items with the same parameters and represent the null hypothesis with no drift. In addition, a random sample of 2000 examinees from the real data is selected without replacement for each administration year, which represents 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 are examined, as outlined in Table 1, for the same set of common items. The assumed dimensionality for each calibration reflects a certain factor structure. First, all 30 items in the three sections are combined together and taken as a one-dimensional model measuring English language ability. Second, because the grammar and vocabulary sections are considered literacy skills in language assessment, while listening is

considered an oracy skill, they are scored as two subscales, but each scale is essentially one-dimensional. For the third condition, the three sections are considered independent from one another and calibrated separately. The sections are truly unidimensional, as they are designed to measure a particular area of English ability. Based on the three-parameter logistic IRT model, the first three calibrations are run using PARSCALE 4.1 (Muraki & Bock, 2003). For comparison, an underlying three-dimensional solution is assumed for model IV. Item parameters are estimated using TESTFACT 4.0. As a result, there are 24 sets (4 calibrations \* 3 years \* 2 data sets) of 400 item parameters estimates after replication.

Table 1. Combinations of Item Calibration

Model	Type <sup>a</sup>	Dimensionality	Item Parameters <sup>b</sup>	Software
I	LGV	One-dimensional	c, a, b	PARSCALE
II	L,GV	Two-dimensional	c, a, b	PARSCALE
III	L,G,V	Three-dimensional	c, a, b	PARSCALE
IV	LGV	Three-dimensional	c, a1, a2, a3, d	TESTFACT

Note: <sup>a</sup> L refers to listening items, G refers to grammar items, and V refers to vocabulary items;  
<sup>b</sup> c refers to guessing parameter that is also known as the lower asymptote parameter, a refers to the item discrimination parameter, b refers to item difficulty parameter,  $\vec{a}$  is the vector of item discrimination parameters, and d is a scalar parameter representing item difficulty.

Because of the indeterminacy of scales and the way different programs are set, item estimates of different calibrations are required to be placed on the same scale for linking. For the unidimensional model, the item parameters fall into a linear relationship, which leads to indeterminacy in the scales of calibrations for different groups of examinees. A linear transformation is 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 resulted in smaller sampling error given other things being equal (Oshima, Raju, & Flowers, 1997). Calibrations of all replications for the first three models were converted to these base scales.

Indeterminacies are issues also raised in scaling calibrated item parameters under multidimensional theory. Three types of indeterminacy are 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 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 is the standard deviation of the observed proficiencies (Li & Lissitz, 2000). The third type of indeterminacy is due to the orientation of the coordinate system. TESTFACT addresses this issue by setting the coordinate system to be orthogonal, which suggests the correlations among the coordinates are set to be zero. Li and Lissitz propose an approach of 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 allows an oblique Procrustes method (Mulaik, 1972) on the basis of work by

Martineau (2004) and Reckase and Martineau (2004). In Reckase (2006), the rotation matrix is defined in that method as:

$$ROT = (\bar{a}_a' \bar{a}_a)^{-1} \bar{a}_a' a_b,$$

where  $\bar{a}_a$  is a  $n \times m$  matrix of discrimination parameters for the reference system of the linked group and  $\bar{a}_b$  is a  $n \times m$  matrix of discrimination parameters for the reference system of the base group. ROT becomes the  $m \times m$  rotation matrix for the discrimination parameters.

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

$$\hat{a}_b = \bar{a}_a ROT$$

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

$$\hat{d}_b = \bar{d}_a + TRAN = \bar{d}_a + a_a (\bar{a}_a' \bar{a}_a)^{-1} \bar{a}_a' (d_b - d_a),$$

where  $d_b$  is a  $n \times 1$  vector of d-parameters for the reference system of the linked group and  $d_a$  is a  $n \times 1$  vector of d-parameters for the reference system of the base group. TRAN becomes the  $m \times 1$  transformation vector for the d-parameters.

### Analysis

Item parameter estimates converted on the same scale across different administration years are compared first for replications of simulated samples and samples from real data. The means of parameter estimates of the 400 replications for the common items are plotted to detect significant discrepancy over time. Such plots can tell the deviation in distributions of the item parameter estimates for real data samples compared to those from simulation samples. The simulation data is assumed to be without any drift in parameter estimates because they were generated using the same set of item parameters. Items that show the most aberrant deviation over time in the real data samples might reveal a drift. Invariant item parameter estimates also suggest a good model/test response data fit. Comparing invariance properties across different calibrations can lead to a more favored model.

Even if differences are observed for parameter estimates of the common items, it is necessary to test whether these differences are statistically significant or are simply due to random error. The standard detection method for differential item functioning (DIF) can 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 Fler (1995), is used here to study IPD. This framework aims to compare test characteristic curves and can 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:

$$P_i(Y_{ij} = 1 | a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i) \frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))},$$

where  $a_i$  is the item discrimination parameter for item  $i$ ,  $b_i$  is the item difficulty parameter for item  $i$ ,  $c_i$  is the lower asymptote parameter for item  $i$ , and  $\theta_j$  is the ability parameter for examinee  $j$ .  $D$  is the scaling constant (1.702) to control for the difference between the logistic function and normal ogive function.

The probability of correctly answering item  $i$  for examinee  $j$  based on the multidimensional three-parameter logistic model (Reckase, 1985; Reckase & Mckinley, 1991) is given by:

$$P_i(Y_{ij} = 1 | \bar{a}_i, c_i, d_i, \bar{\theta}_j) = c_i + (1 - c_i) \frac{\exp(D\bar{a}_i'\bar{\theta}_j + d_i)}{1 + \exp(D\bar{a}_i'\bar{\theta}_j + d_i)},$$

where  $\bar{a}_i$  is an  $m \times 1$  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  $i$ ,  $\bar{\theta}_j$  is an  $m \times 1$  vector of the ability parameters for examinee  $j$ ,  $m$  refers to the number of dimensions for ability parameters, and  $D$  is the scaling constant (.702).

IRT-based true scores are estimated as:

$$\tau_j = \sum_{i=1}^k P(Y_{ij} = 1 | a_i, b_i, c_i, \theta_i) \text{ for the unidimensional model and}$$

$$\tau_j = \sum_{i=1}^k P(Y_{ij} = 1 | \bar{a}_i, d_i, c_i, \bar{\theta}_i) \text{ for the multidimensional model.}$$

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:

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

where  $E$  is the expectation taken over either the reference group or the focal group,  $\mu$  and  $\sigma$  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):

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

where  $P_{iF}$  and  $P_{iR}$  are the probability of a correct response at a given theta value (or vector for multidimensional model) using item parameter estimates 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 true score differences 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  $(1 - \alpha)$  percentile with the type I error rate of  $\alpha$ . Given the choice of  $\alpha$  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 is that the mean of the NCDIF distributions from simulation samples is equivalent to that from the real data samples. The

larger the number is, the more frequent the NCDIF values in the real data sample are rejected as being the same distribution as the null.

## Results

### Descriptive Statistics

The descriptive statistics shown in Table 2 include means and standard deviations for scale scores for items in each section, which are based on the number of correct responses. The number-correct score means of these items for examinees at Year 6 are consistently lower than the previous two years. Kuder-Richardson Formula 20 (K-R20) estimates of reliability, known as a special case of Cronbach's alpha, are used for ordinal dichotomies in particular. That is, the items are 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 is lower than the criteria value of 0.9 for a homogenous test. The K-R20 is known as 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) is included in this study. Also reported is the number of examinees who were administered the test each year.

Table 2. Descriptive Statistics and Scale Reliability

	Case	Total			Listening		Grammar		Vocabulary	
		K-R20	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Year 1	17151	0.739	20.38	4.54	7.65	1.88	7.09	2.06	5.65	2.11
Year 3	22099	0.717	20.81	4.31	7.73	1.78	7.08	2.02	6.01	2.01
Year 6	33027	0.742	19.86	4.62	7.58	1.89	6.75	2.07	5.54	2.15
Total	72277	0.735	20.28	4.53	7.64	1.85	6.93	2.06	5.71	2.11

Note: S.D.: Standard deviation; K-R20: Kuder-Richardson Formula 20

### Dimensionality

The eigenvalues from the tetrachoric correlation matrix were computed by TESTFACT, shown in Table 3. There is no output of eigenvalues for NOHARM because the program analyzes the sample proportion correct for item pairs instead of a tetrachoric correlation matrix. The first three factors all have eigenvalues that are greater than one. The first factor is dominant with an eigenvalue around seven, and the second factor is strong with an eigenvalue close to three. The ratios of factor differences were also computed and are presented in the last column of Table 3. The datasets meet the criteria except for Year 3, but the FDRI value is very close to three, which confirms that the first factor dominates while the other factors from the second onward make a relatively minor contribution.

As would be expected, the eigenvalue analysis verified the existence of a dominant factor and two minor factors for the test data. The first extracted factor approximately accounted for 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 attributed less than 4% of the variance. Reckase (1979) suggests the first factor accounting for at least 20% of total variance verifies a dominant underlying latent factor for items concerned. The percentages associated with the first factor were close to critical values, implying an approximation to unidimensionality for the test data.

Table 3. Eigen Values, FDRI, and Percentage of Variance Explained using TESTFACT

TESTFACT	Eigen Values			FDRI	Percentage of Variance Explained		
	F1	F2	F3		F1	F2	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%

Residuals and fit statistics are also compared for the three factor models and are summarized in Table 4. A hypothesized one-factor model, resulting in RMSRs that are close to 0.09, does not provide support for a good fit to the data. However, the values decrease substantially to approximately 0.05 for the two-factor model, and around 0.03 for the three-factor model. Comparison of these results with those for the one-factor model suggests that the three-factor solution provides the best fit to the data.

The RMSRs are also available in NOHARM, but the residual statistic is based on the difference between the observed and model-based proportions of correct response for each pair of items. The RMSRs are generally less than 0.01 and a gradual decrease is observed across the three solutions. The Tanaka indexes are also generally greater than the criteria value of 0.95, indicating a good model fit. Index values for the three-factor model solution are close to one, which suggests a nearly perfect fit to the data. The appreciable improvement of model fit occurs after adding the second and the third factors.

Table 4. Goodness-of-Fit Statistics for TESTFACT and NOHARM Exploratory Solutions

TESTFACT	Root Mean Square Error of Approximation (RMSEA)			Root Mean Square of Residuals (RMSR) <sup>a</sup>		
	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

NOHARM	Root Mean Square of Residuals (RMSR) <sup>b</sup>			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

<sup>a</sup> RMSR in TESTFACT are based on the residual correlations as the difference between model-based and observed item correlations.

<sup>b</sup> RMSR in NOHARM are based on the residual correlations as the difference between model-based and observed proportions of correct responses.

Even though the exploratory analysis evidences a dominant first factor that explains most of the variability in observed scores, it is necessary to check whether the other two minor factors are essentially different than the first factor and are significant measurable constructs. In addition to the eigenvalue and residual analyses shown above, graphic tools for measures of fit are also included to compare models of different orders and to provide substantive support for multidimensionality.

Figure 1 displays the item vector plots for test data each year and the test data with three years combined. These item vectors are plotted with regard to orthogonal solutions with the three-factor model. The vector plots reveal the separation of items into more than one group. The listening items (in red) for all four plots show a clear pattern of pointing in the same direction, which suggests they constitute an essentially unidimensional scale. However, the grammar items (in blue) and vocabulary items (in black) vary widely in their orientations.

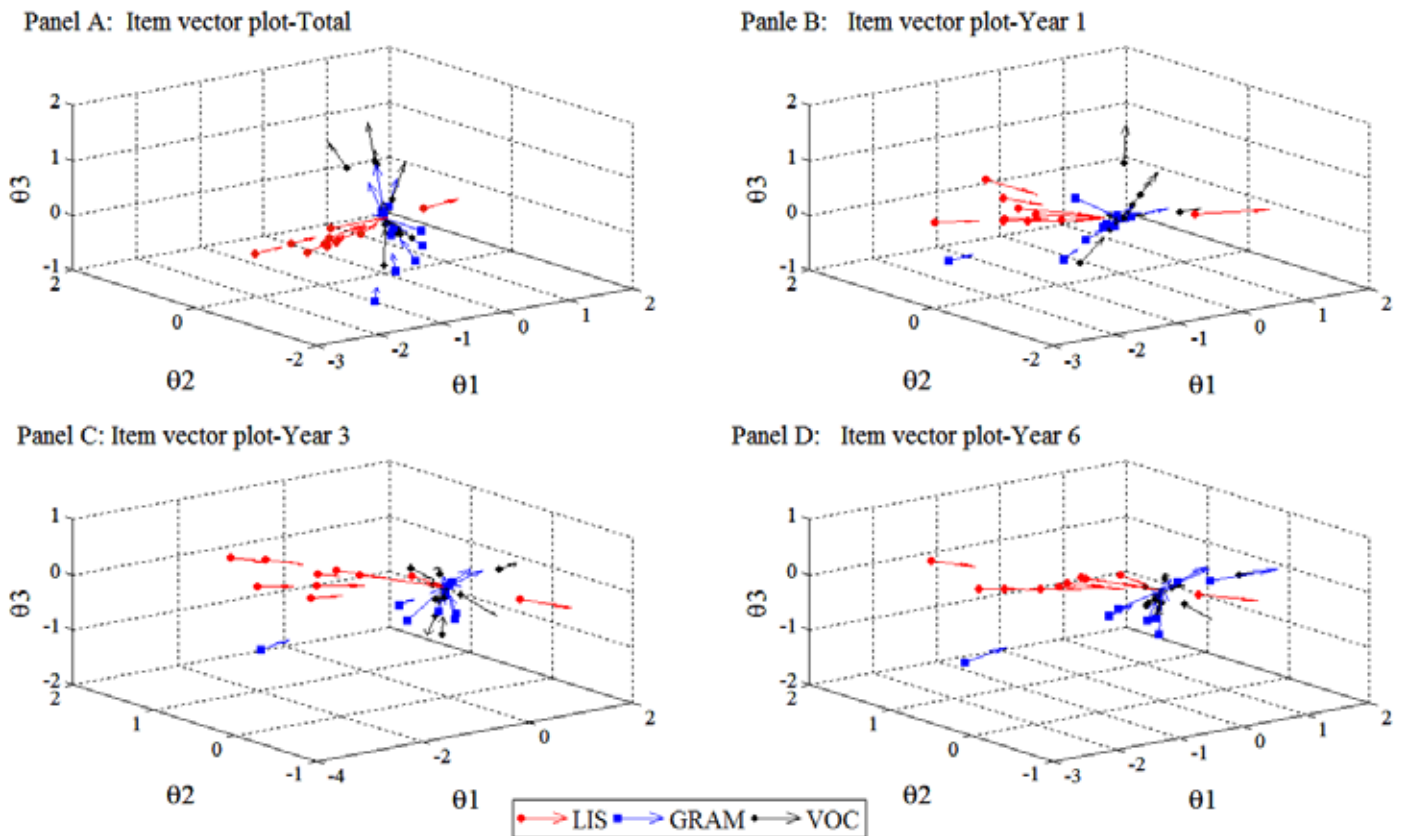


Figure 1. Item Vector Plots for Total, Year 1, Year 3, and Year 6

In panel A both grammar and vocabulary items mix with each other but appear to comprise two clusters. For panels B and D, both types of items mostly span a common vector. In panel C, most of the items are oriented in two different directions but within a small degree. In general, the vector plots of all 30 linking items result in three sets of items, each measuring

essentially the same composites of skills. The listening items measure the same combination of skills and the other 20 items measure two different composites of abilities.

It is important that the factor solutions are interpretable for the purpose of multidimensionality (Gorsuch, 1983). Each of the items had substantial loadings on one factor. As a result, these items can be used as indicators to represent the factors from a three-factor solution. Inspection of Promax rotated factor loadings, given in Table 5, show that the three factors extracted are mathematically acceptable and nontrivial. The highest loadings are highlighted in bold. Only listening items load high on the second factor for each year, and that can be identified as the factor for listening capabilities. Most of the grammar items load on the first factor. The vocabulary items are separated into two groups, six of which load on the first factor, and four on the third factor. Though grammar and vocabulary items do not cluster exactly as designed, the patterns for the factor loadings are similar across administrations.

The interfactor correlations for the Promax rotated solution are given in Table 6. The highest correlations are greater than 0.60, between the first factor and the third factor, represented by all grammar and vocabulary items. This is consistent with the test design that assumes both items measure English literacy skills. These two factors are also moderately correlated with the second factor, indicating oracy skills, represented by all listening items.

Table 5. Promax Rotated Factor Loadings Based on Three-Factor Model

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

	Total			Year 1			Year 3			Year 6		
	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
V25	<b>0.38</b>	0.16	0.34	<b>0.41</b>	0.15	0.29	<b>0.35</b>	0.16	0.31	0.37	0.17	<b>0.41</b>
V26	<b>0.46</b>	0.07	0.09	<b>0.39</b>	0.08	0.13	<b>0.43</b>	0.07	0.09	<b>0.49</b>	0.05	0.09
V27	<b>0.45</b>	0.02	0.06	<b>0.57</b>	0.01	0.10	<b>0.48</b>	0.02	0.08	<b>0.34</b>	0.02	0.03
V28	<b>0.57</b>	0.01	0.03	<b>0.51</b>	0.03	0.11	<b>0.51</b>	0.02	0.07	<b>0.58</b>	0.03	0.04
V29	<b>0.63</b>	0.11	0.03	<b>0.66</b>	0.10	0.03	<b>0.57</b>	0.09	0.01	<b>0.65</b>	0.16	0.01
V30	<b>0.29</b>	0.13	0.12	<b>0.36</b>	0.10	0.07	<b>0.30</b>	0.13	0.08	<b>0.25</b>	0.15	0.16

Table 6. Promax Factor Correlations Based on Three-Factor Model

	Total			Year 1			Year 3			Year 6		
	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	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

In summary, a similar number of dimensions were identified by eigenvalues analyses, residual and fit statistics, and graphic assessment. A three-factor model solution is generally supported, while there is also evidence of a dominant first factor or general factor. Although residual statistics suggest a parsimonious one-factor solution could fit the model well, graphic analyses and factor loadings imply a model with at least two or three factors should be considered for better identification and interpretation. Besides the second factor distinctively represented by listening items, three-fourths of the grammar and vocabulary items have the highest loadings on the first factor and one-fourth on the third factor. The third factor, much weaker than the other two, shows some evidence of representing uniqueness of vocabulary items that could not be compensated by the grammar items. These tests were in essence three-dimensional components that were in agreement with the theoretical structure in terms of English proficiency in three skill areas.

### Invariance 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 7. Guessing parameters ( $c$ , in the first column) were calibrated with BILOG-MG from 72,277 real data samples with the three administrations combined. Also included are the vectors of  $a$  and  $d$  parameters calibrated by NOHARM under a three-dimensional compensatory logistic model. Items with the highest loadings are used to define the axes, which were rearranged for calibration with item 8 placed at the first place. Item 18 and item 23 follow as the second and the third items. By default, the  $a$  parameters for the first item (L08) are set to zero for  $a_2$  and  $a_3$ , and the second item (G18) to zero for  $a_3$ .

These item parameters were fixed and submitted to TESTFACT for estimating the ability distribution for each year of administration. As shown in Table 8, the means are close to zero but there are differences in the mean levels in terms of the three constructs across years. The examinees who took the test in the third year in general have the highest mean ability levels (all equal to 0.1), while those in the sixth year are relatively low. Item responses

for the test in the first year resulted in a correlation of 0.46 between  $\theta_1$  and  $\theta_2$ , 0.42 between  $\theta_1$  and  $\theta_3$ , and 0.64 between  $\theta_2$  and  $\theta_3$ . For the third year, the correlations were 0.48 between  $\theta_1$  and  $\theta_2$ , 0.40 between  $\theta_1$  and  $\theta_3$ , and 0.64 between  $\theta_2$  and  $\theta_3$ . The test data in the third year show that  $\theta_1$  and  $\theta_2$  are correlated at 0.53,  $\theta_1$  and  $\theta_3$  are correlated at 0.43, and  $\theta_2$  and  $\theta_3$  are correlated at 0.62. These correlations are corrected for attenuation and modeled for the error-free measures of the constructs, which should be higher than the observed correlations. The variance-covariance matrices were then computed and are presented in Table 8.

The means and variance-covariance matrices were used for generating item response data for each year of administration. The same set of item parameter estimates in Table 7 was input into the three-dimensional extension of the three-parameter logistic compensatory model for data generation. As a result, the simulated test data assumes the item parameter estimates are equivalent except for measurement error. At the same time, samples were randomly selected from real data for each year of administration. Estimates of item parameters were calibrated on the basis of the item response data from these samples. The results are presented in the following section for comparing estimates from real data to those from simulation data assuming different models.

Table 7. Parameter Estimates for Linking Items used for Simulation

Item	c	a1	a2	a3	d
L01	0.267	<b>0.585</b>	0.032	0.041	1.270
L02	0.229	<b>0.575</b>	0.071	0.011	0.979
L03	0.344	<b>0.733</b>	0.005	0.064	-0.458
L04	0.079	<b>0.532</b>	0.129	0.037	0.697
L05	0.270	<b>0.879</b>	0.248	0.079	1.008
L06	0.125	<b>0.452</b>	0.086	0.081	0.538
L07	0.076	<b>0.603</b>	0.134	0.069	0.990
L08 *	0.384	<b>1.186</b>	0.000	0.000	1.009
L09	0.168	<b>0.802</b>	0.036	0.116	0.749
L10	0.000	<b>0.419</b>	0.087	0.132	0.229
G11	0.119	0.317	<b>0.567</b>	0.331	-0.121
G12	0.101	0.135	<b>0.398</b>	0.160	0.112
G13	0.063	0.087	<b>0.619</b>	0.185	0.501
G14	0.328	0.146	<b>0.456</b>	0.265	0.501
G15	0.140	0.405	<b>0.388</b>	0.100	1.542
G16	0.000	0.596	<b>0.745</b>	0.006	0.724
G17	0.500	0.350	0.524	<b>0.784</b>	-0.103
G18 *	0.023	0.030	<b>0.971</b>	0.000	0.590
G19	0.284	<b>0.591</b>	0.466	0.292	-0.254
G20	0.190	0.234	<b>0.317</b>	0.224	0.514
V21	0.200	0.468	0.152	<b>0.908</b>	-0.351
V22	0.500	0.437	0.441	<b>0.857</b>	0.862
V23 *	0.200	0.026	0.209	<b>1.028</b>	-1.055
V24	0.315	0.283	0.357	<b>0.489</b>	0.078
V25	0.255	0.062	<b>0.555</b>	0.520	-0.747
V26	0.121	0.123	<b>0.518</b>	0.239	0.213

Item	c	a1	a2	a3	d
V27	0.125	0.167	<b>0.454</b>	0.083	0.180
V28	0.071	0.258	<b>0.669</b>	0.246	0.291
V29	0.015	0.109	<b>0.681</b>	0.190	0.405
V30	0.016	0.294	<b>0.362</b>	0.250	-0.626

\* Items used to anchor the axes.

Table 8. Parameter Estimates for Linking Items used for Simulation

Item	Year 1			Year 3			Year 6		
	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

### Results of Model I

The plots summarizing the means of the item parameter estimates over years are displayed for Model I in Figure 2. This model assumes a dominant one-factor solution for all 30 items. These item parameter estimates are linked to the same scale. The reference scale is the parameter estimates calibrated on all the data with three administrations combined.

The panels on the left exhibit the means for parameter estimates from 400 simulation samples. The upper panel compares estimates for  $a$  parameters and the bottom panel contrasts estimates for the  $b$  parameter. As expected, the means are fairly similar across years of administration and nearly fall on the same line, which suggests the items have invariant difficulty values and discriminate equally well over time. A few exceptions are item 17 and item 22 in terms of the item discrimination.

The panels on the right exhibit the means for parameter estimates from 400 samples from the real data. The upper panel compares estimates for  $a$  parameters and the bottom panel contrasts estimates for the  $b$  parameter. Compared to the simulation data, the means mostly fall on a single line but have a clear variation across administrations. Estimates of item difficulty remain stable over years except for those of item 3 and item 29. The item discrimination power is relatively variable, especially for that of items 5, 17, 21, and 22.

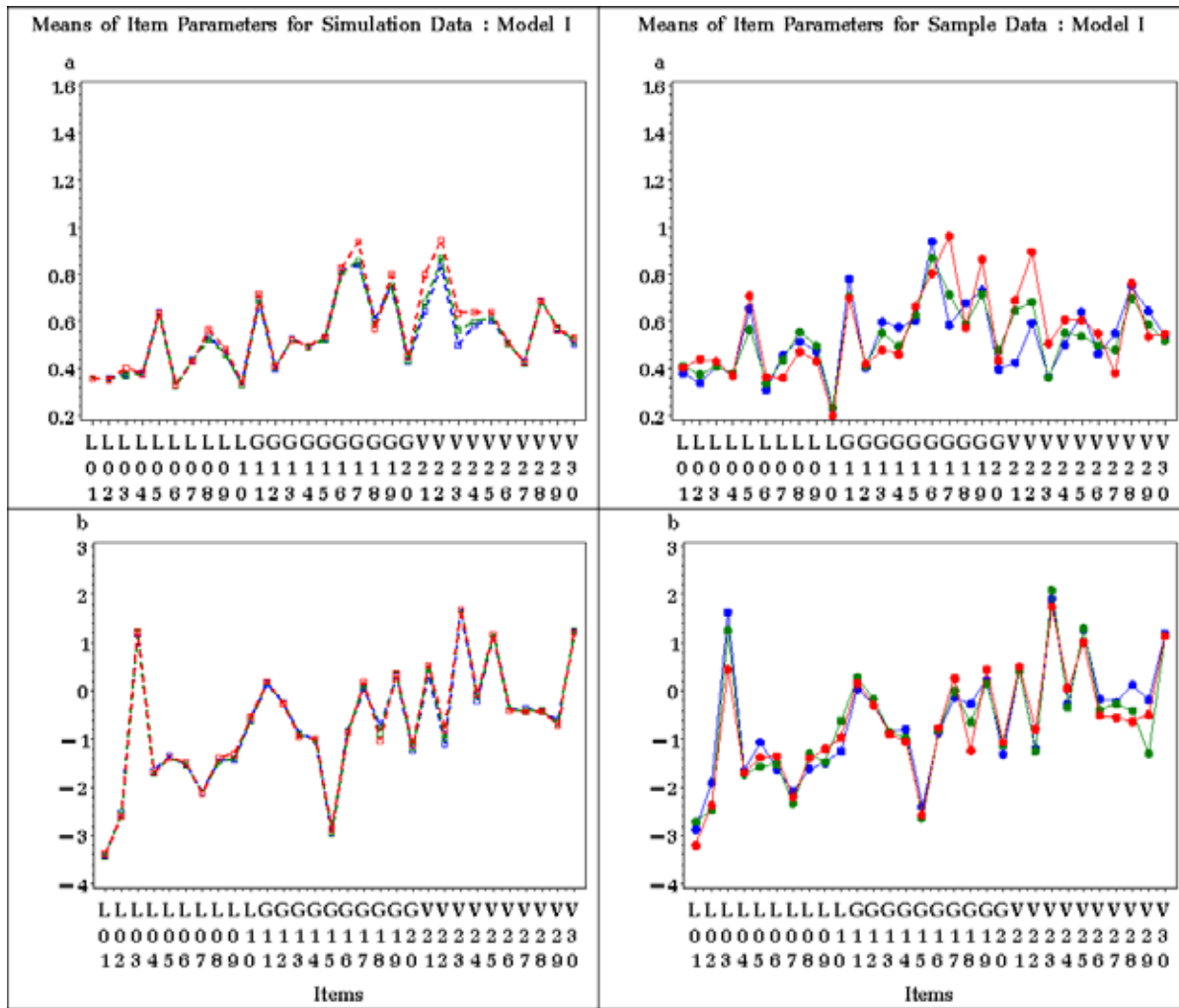


Figure 2. Mean plots of item parameter estimates for simulation data and real data assuming Model I<sup>1</sup>

### Results of Model II

The plots summarizing the means of the item parameter estimates over years are displayed for Model II in Figure 3. This model assumes two different test scales for all 30 items with one indicating oracy skills and the other indicating literacy skills. These item parameter estimates are linked to the same scale. The reference scale is the parameter estimates calibrated on all the data with three administrations combined.

The panels on the left exhibit the means for parameter estimates from 400 simulation samples. The upper panel compares estimates for  $a$  parameters and the bottom panel contrasts estimates for the  $b$  parameter. As expected, the means are fairly similar across years of administration and nearly fall on the same line, which suggests the items have invariant

<sup>1</sup> The dotted lines represent the parameter estimates based on the 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, those in green represent item estimates from Year 3, and those in red represent item estimates from Year 6.

difficulty values and discriminate equally well over time. A few exceptions are item 17 and item 22 in terms of item discrimination.

The panels on the right exhibit the means for parameter estimates from 400 samples from the real data. The upper panel compares estimates for  $a$  parameters and the bottom panel contrasts estimates for the  $b$  parameter. Compared to the simulation data, the means almost fall on a line but have a clear variation across administrations. Estimates of item difficulty remain stable over years except for those of items 3 and 29. The item discrimination power are relatively variable, especially those of items 3, 17, 21, and 22.

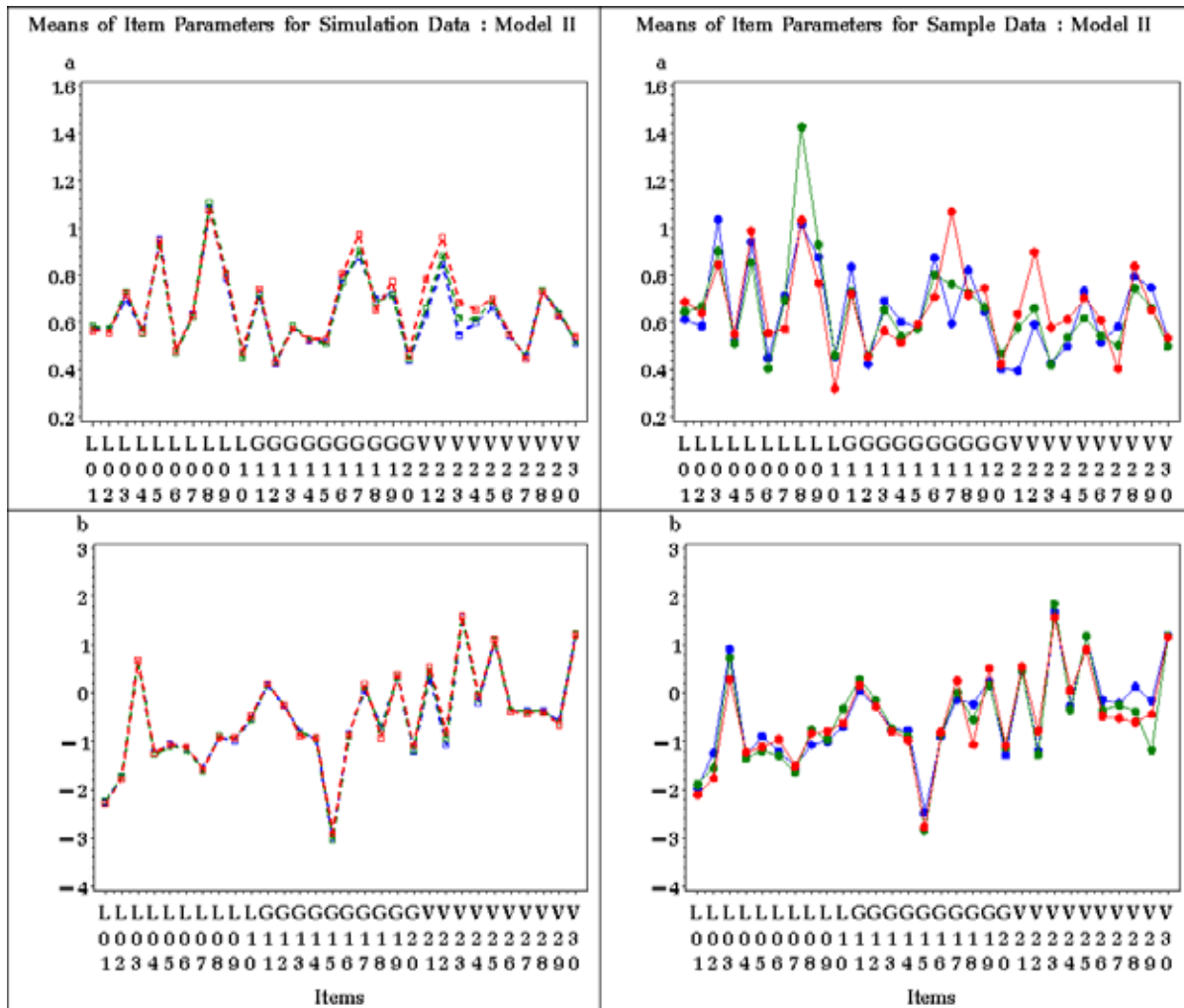


Figure 3. Mean plots of item parameter estimates for simulation data and real data assuming Model II (See note 1 above.)

### Results of Model III

The plots summarizing the means of the item parameter estimates over years for Model III are displayed in Figure 4. This model assumes each set of items in the listening, grammar, and vocabulary sections represents a test subscale. These item parameter estimates are linked to the same scale. The reference scale is the parameter estimates calibrated on all the data with three administrations combined.

The panels on the left exhibit the means for parameter estimates from 400 simulation samples. The upper panel compares estimates for  $a$  parameters and the bottom panel contrasts estimates for the  $b$  parameter. As expected, the means are fairly similar across years of administration and nearly fall on the same line, which suggests the items have invariant difficulty values and discriminate equally well over time. An exception is item 17 in terms of the item discrimination.

The panels on the right exhibit the means for parameter estimates from 400 samples from the real data. The upper panel compares estimates for  $a$  parameters and the bottom panel contrasts estimates for the  $b$  parameter. Compared to the simulation data, the means almost fall on a single line but have a clear variation across administrations. Estimates of item difficulty remain stable over years except for those of items 18 and 29. The item discrimination power is relatively variable, especially those for items 8, 17, 21, and 22.

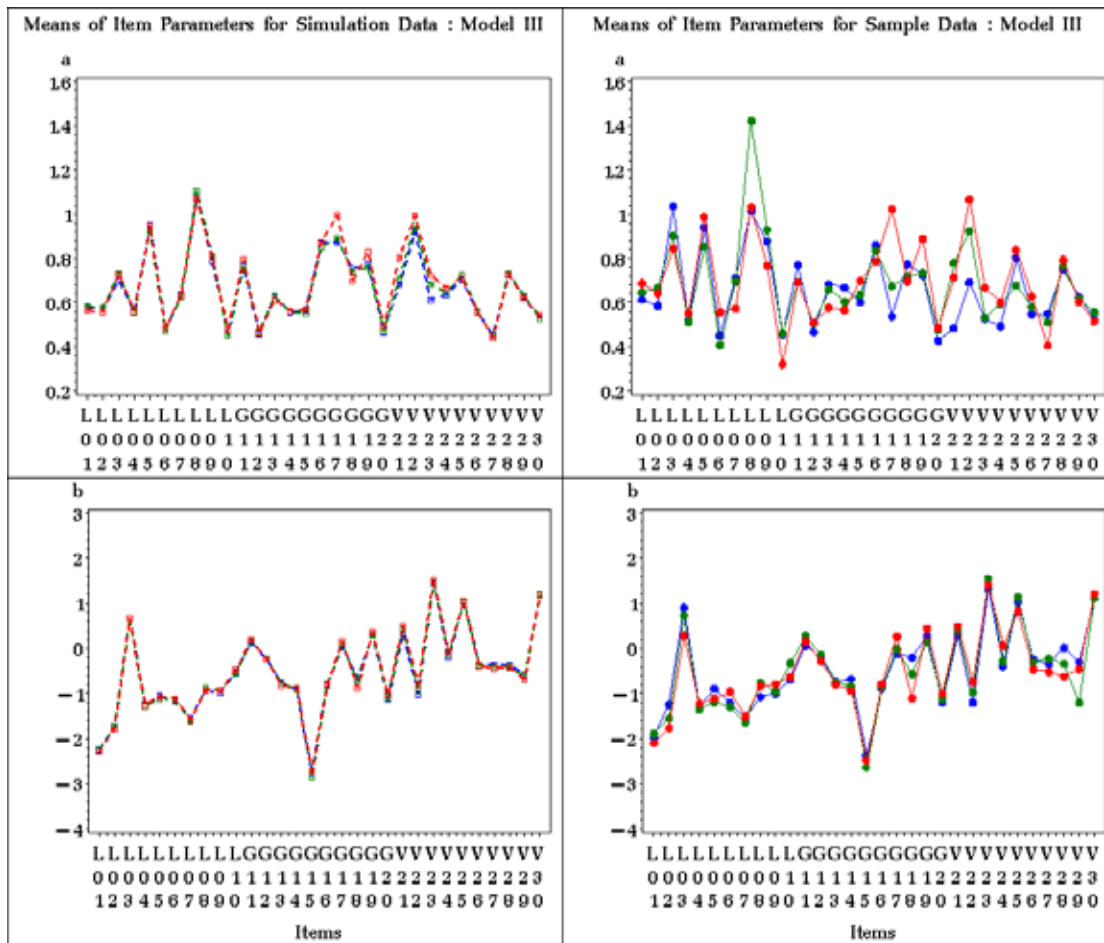


Figure 4. Mean plots of item parameter estimates for simulation data and real data assuming Model III (See note 1 above.)

### Results of Model IV

The plots summarizing the means of the item parameter estimates over years for Model IV are displayed in Figure 5 and Figure 6. This model assumes a three-dimensional extension of three-parameter compensatory models resulting in three skill areas. These item parameter estimates are linked to the same scale. The reference scale is the parameter estimates calibrated on all the data with three administrations combined.

The panels on the left exhibit the means for parameter estimates from 400 simulation samples. The upper panel in Figure 5 compares estimates for  $a1$  parameters and the bottom panel contrasts estimates for  $a2$  parameters. The upper panel in Figure 6 compares estimates for  $a3$  parameters and the bottom panel contrasts estimates for  $d$  parameters. As expected, the means are approximately equivalent across years of administration and overlap on the same line, which suggests the items have invariant difficulty values and discriminate equally well over time. One exception was item 18 in terms of the  $a2$  parameter.

The panels on the right exhibit the means for parameters estimates from 400 samples from the real data. The upper panel in Figure 5 compares estimates for  $a1$  parameters and the bottom panel contrasts estimates for  $a2$  parameters. The upper panel in Figure 6 compares estimates for  $a3$  parameters and the bottom panel contrasts estimates for  $d$  parameters. Compared to the simulation data, the means mostly fall on a line but have a slight variation across administrations. Estimates of item difficulty remain stable over years except for that of item 17. The item discrimination powers are relatively variable, especially for those of items 3 and 8.

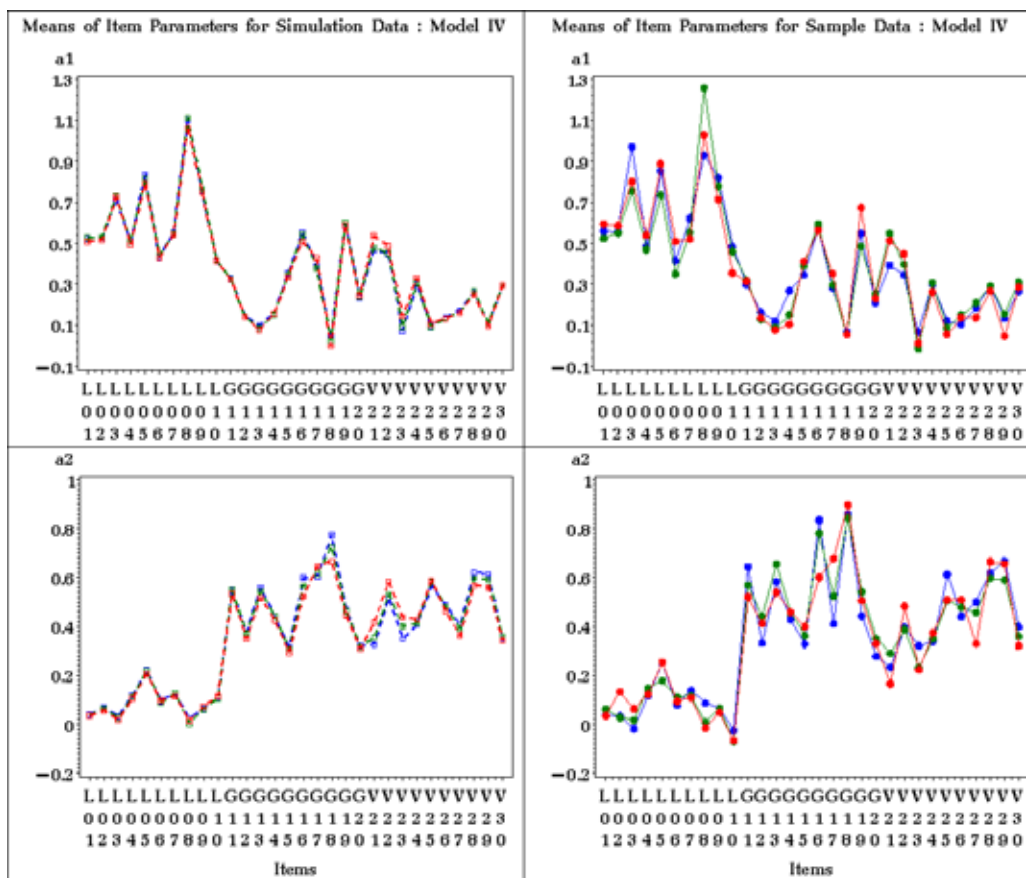


Figure 5. Mean plots of estimates for  $a1$  and  $a2$  for simulation data and real data assuming Model IV (See note 1 above.)

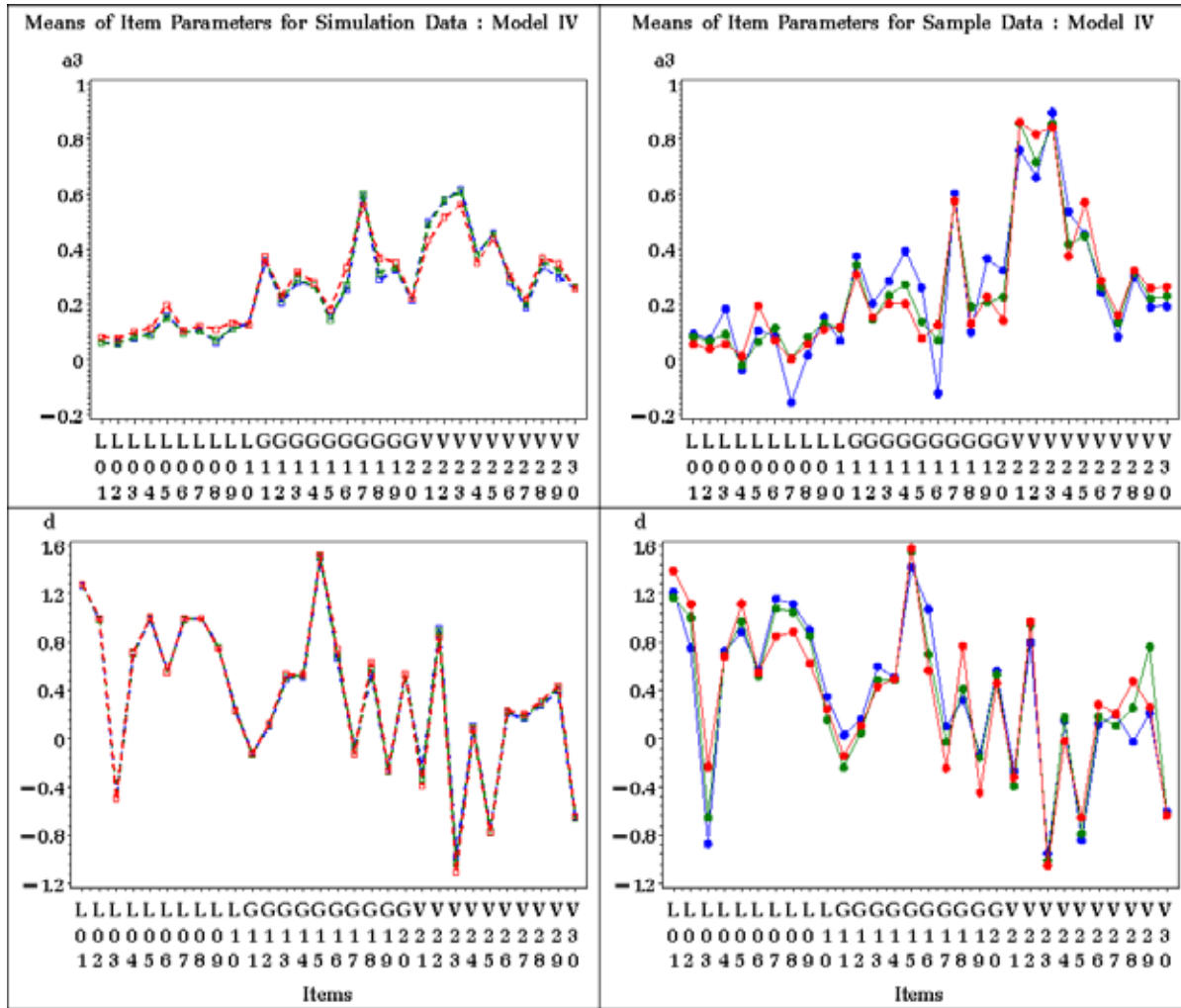


Figure 6. Mean plots of estimates for  $a_3$  and  $d$  for simulation data and real data assuming Model IV (See note 1 above.)

The counts of NCDIF index values that are greater than the cut-off values set are summarized in Tables 9, 10, and 11. Each table represents the results from one year of administration. Out of the 400 replications, the numbers suggest the deviation of the distribution of NCDIF for real data from the distribution based on simulation data.

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

For the first year, the largest number in rejection areas was 400 times (100%) for items 2, 18, 28, and 29 under Model I. The frequency tends to decrease as the number of factors increases in the models. In general, the overall frequency of being rejected has a pattern of gradual decline. Model IV has only a few cases being rejected. The  $\alpha = 0.01$  case shows no sample NCDIF being rejected except for one item: item 20.

Table 9. Number of Replications with NCDIF above the Cut-Value for First Year

Item	Model I		Model II		Model III		Model IV	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
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

For the third year, the largest number of rejections is 400 times for item 29 under Model I. The frequency tends to decrease as the number of factors increases in the models. In general, the overall frequency of being rejected has a pattern of gradual decline. Underlying MIRT models, Model IV has only a few cases being rejected. The  $\alpha = 0.01$  level has only one case being rejected for most of the items.

Table 10. Number of Replications with NCDIF above the Cut-Value for Third Year

Item	Model I		Model II		Model III		Model IV	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
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

For the sixth year, the largest number of rejections is 398 times (99.5%) for item 28 under Model I, followed by item 18 with 389 times. The frequency tends to decrease as the number of factors increases in the models. In general, the overall frequency of being rejected decreases gradually. Model IV has only a few cases being rejected. The  $\alpha = 0.05$  level shows five items being rejected a few times but no item is rejected for the NCDIF index at the significance level of 0.01.

Table 11. Number of Replications with NCDIF above the Cut-Value for Sixth Year

Item	Model I		Model II		Model III		Model IV	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
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

## Discussion

This study empirically examined the effect of multidimensionality upon the invariance property of item parameter estimates in an IRT model. The ECPE data were used to investigate the potential drift of both item difficulty and item discrimination estimates for the same set of items. Only these 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 are sensitive to multiple dimensions. Samples selected from real data were compared to the simulation data generated under multidimensional model.

Multiple dimensions were identifiable for the 30 common items used in the ECPE even when traditional methods using eigenvalue analyses identified a single dimension. The results of residual analyses and item vector plots suggest that three dimensions are optimal solutions. The dimensionality might be underestimated by conventional techniques because they rely on a dominant dimension and shared variance. The multiple constructs are highly correlated but they measure different composites of English proficiency.

Preliminary analyses comparing mean plots show that the item parameter estimates for simulated samples remain stable as expected. Especially for item difficulty estimates, the means are equivalent over time. There are a few items that exhibit slight deviation for item discrimination values, which was attributed to measurement error. Consistent stability was also observed for different models.

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

In general, the item difficulty indices exhibit a very high degree of invariance across samples, even for calibrations in the one-dimensional model. No obvious negative effect of multidimensionality on the invariance property was observed. The models with lower dimensions show a tendency of having slightly less invariance estimates than the models with higher dimensions. The estimation of item difficulty parameters is robust and remains stable for both unidimensional and multidimensional models.

The item discrimination parameter estimates are generally less invariant than the item difficulty values. The degree of invariance of item discrimination parameter estimates also increases steadily as the dimensions of models increase, implying that the IRT discrimination parameter estimates do not maintain a high degree of invariance for items sensitive to multiple dimensions. In addition, the number of items with dissimilar discrimination estimates decreases 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 are compared. It shows that the differences are not due to random noise but lead to different item characteristic curves. The count of NCDIF values greater than the cut-off values provides guidance for what degree of parameter variation is within acceptable limits.

As a result, the analyses of real test data presented as examples in this paper show that there is evidence of an effect of multidimensionality on parameter invariance. Multidimensional models generally exhibit less variation than unidimensional models, even for models assuming three dimensions based on sections. The results show that the choice of models for calibration and linking tend 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 dimensionality addressed. For items that are sensitive to multiple dimensions, models with higher dimensions produce similar indices across forms and are consistently the best among the models. The observed IPD using unidimensional models might indicate that inadequate dimensionality was addressed.

## **Implications and Limitations**

The findings of this study have important implications for the ECPE and other long-term large-scale assessments (e.g., the Examination for the Certificate of Competency in English). 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 eigen values or scree plots. However, this assumption cannot be completely met by any set of test data. The conventional exploratory factor analysis assessment might be misleading, especially for test data with highly correlated factors. The findings of other studies are actually based upon a combination of measures that is the aggregate of multiple measures. The results are likely to mask what should be differential results related to invariance of item parameters. Researchers and parishioners should be cautious against assuming unidimensionality and the property of parameter invariance might be misleading with 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 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 assessment.

The results must be considered in light of study limitations. The linking items were only a small part of the whole set of test items. Analysis using only linking items is likely to lead to a source of additional sampling error. This study illustrates to some extent the robustness of item response theory under the violation of the assumption of unidimensionality. Though evidence of drift in the item parameters was found, other factors might confound the sources of IPD. For instance, some items might have a large magnitude of IPD if they are administered in two different locations on a test, especially for 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 is at what point this might lead to measurement nonequivalence and cause error in linking and equating. The crucial question is at what point does parameter variation become 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 might result in stronger prediction of the sources of drift in item parameters.

## **Acknowledgments**

The author thanks the University of Michigan English Language Institute (ELI-UM) Spaan Fellowship Committee for permission to use data and for providing funding for this research. The author also thanks Dr. Jeffrey Johnson and Eric Lagergren for editing and reviewing. Tian Song at ELI-UM helped provide test booklets.

## References

- Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. *Applied Psychological Measurement in Education*, 1(3), 215–222.
- 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.
- 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., & Pfeifferberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25(4), 275–285.
- 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.
- 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.
- DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, 17(3), 265–300.
- Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22(1), 33–51.
- 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(4), 369–377.
- Gorsuch, R. L. (1983). *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 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.
- 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. (1985) Methodology Review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(12), 139–164.
- 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(1), 141–151.
- 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.
- Li, Y. H., & Lissitz, R. W. (2000). An Evaluation of the Accuracy of Multidimensional IRT Linking. *Applied Psychological Measurement*, 24(2), 115–138.

- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Martineau, J. A. (2004). *The Effects of Construct Shift on Growth and Accountability Models*. Unpublished doctoral dissertation, Michigan State University, East Lansing.
- McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30(1), 23–40.
- 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.
- 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: Scientific Software International.
- Muthén, L. K., & Muthén, B. O. (2001). *Mplus user's guide (version 2)*. Los Angeles: Muthén & Muthén.
- Oshima, T. C. (1994). The effects of speededness on parameter estimation in item response theory models. *Journal of Educational Measurement*, 31(3), 200–219.
- 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(3), 253–272.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, 4(3), 207–230.
- Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement*, 9(4), 401–412.
- Reckase, M. D. (2006). Multidimensional Item Response Theory. In C.R. Rao, S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics*, 26, North Holland: Elsevier.
- 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(4), 361–373.
- 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(4), 353–368.
- Stocking, M. L. & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, 7(2), 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(2), 125–141.
- Sykes, R. C., & Fitzpatrick, A. R. (1992). The Stability of IRT b Values. *Journal of Educational Measurement*, 29(3), 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.
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The Effect of Item Parameter Drift on Examinee Ability Estimates. *Applied Psychological Measurement*, 26(1), 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: Scientific Software International.
- 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.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 for Windows, Chicago: Scientific Software International.