Multidimensional IRT models for Composite Scores

Shu Jing Yen
Leah, Walker

Paper Presented at the 2007 Annual Meeting of the National Council of Measurement in Education, Chicago

ABSTRACT

Tests of English Language Proficiency are often designed such that each section of the test measures a single latent ability. For instance an English Proficiency Assessment might consist of sections measuring Speaking, Listening, and Reading ability. However, Overall English Proficiency and composite abilities are naturally multidimensional. This multidimensionality may even include the case where each item measures more than one latent ability. Parameters for these examples can be estimated within a one-dimensional or multidimensional item response modeling framework. This study compares two common strategies for composite scoring in Test of English Language Proficiency to composite scoring strategies derived from multidimensional models. An example is taken from a sample of 12,008 elementary school students who participated in a State English Proficiency Assessment. Factor analytic techniques were performed to verify the dimensional structure of the composite domains for the Oral scale. Item Response Models with separate unidimensional calibration, simultaneous unidmensional calibration, and multidimensional calibration were examined. Model-data fit, information functions, ability estimates, and interpretive capability among the three different types of composite scores were examined.

INTRODUCTION

The motivation for No Child Left Behind is to ensure quality education for all students. This includes students who are still learning English. While English Proficiency examinations programs were in existence before passage of NCLB, after its passage new testing and reporting requirements were enacted. One of these requirements is that states assess English Language learners in English Proficiency until they are deemed proficient. Their proficiency is determined by scores in the required reporting areas. These are Listening, Speaking, Reading, and Writing, and the Composite areas of Comprehension, Oral Skills, and Overall English Language Proficiency. While the scores are required, the method of computing these scores is largely left to the states and their contractors.

The most common strategy for producing the Listening, Speaking, Reading, and Writing scores is to administer four individually calibrated subtests and to report a score on each. For the composite areas, various methods exist. Some methods include taking various linear combinations of scores on the subtests. This may be a simple average, a reliability weighted linear combination, or an implicit weighting using discrimination parameters as weights. Since any scoring system to be used for such important decisions should be designed to produce reliable and valid scores, the item response modeling approach is preferred although not generally employed.

Within the item response modeling framework, both one-dimensional and multidimensional models can be used to parameterize composite domains. However since the most common practice in designing such tests is to ensure that each subtest is one-dimensional, it is not necessarily guaranteed that the composite domains which consist of more than one subtest are one-dimensional. Thus it is likely that multidimensionality exists in the composite domains (at least). When multidimensionality exists, multidimensional item response theory has been shown to produce more accurate and efficient parameter estimates (Reckase, 1985). Thus, the use of multidimensional item response theory in composite score creation may provide better composite estimates.

In many achievement-testing situations it is useful or sometime required to combine measures from different components into a single score. This may involve combining scores from measures of a single content area with different item types or combining measures of different content areas with the same or different item types. One example is to combine answers from essay questions and the multiple choice questions on a writing assessment (Wainer & Thissen, 1993). In this example, the goal of the assessment is to create some uni-dimensional writing score. Another example is to combine a quantitative subtest and a verbal subtest to create a composite score. In this example, scores on distinct subtests are combined to form a composite score, even though the subtests are designed to measure different competencies. So the goal of this example is create an overall proficiency score combining information from multiple content areas. The most common way to combine the component scores is to take their weighted sum or average. Assuming that the final score is to be a weighted sum of the separate subtests, the question is reduced to the choice of the weight to be assigned to each subtest.

A wide variety of approaches to weighting items and subtests have been described. Gulliksen (1950) provided an extended discussion of early research on weighting. Gulliksen suggested two possible methods: to weight the components in order to maximize the reliability of the composite, or to reply solely on expert judgments.

Wainer and Thissen (1993) examined these strategies in the context of Advanced Placement tests. It was suggested that since the goal is to enhance the overall reliability of the test; the most efficient way to achieve this goal is to weight subcomponents by their reliabilities. Under this strategy, more reliable components are given more weight. Later Rudner (2001) continued the discussion by showing that validity can decrease as the reliability is increased by giving more weight to the more reliable but less valid component. Thus, it was shown that there can be a tradeoff between reliability and validity under the reliability weighting strategy. Kane and Case (2004) make a similar argument using simulation study. They showed that giving extra weight to the more reliable of the two observed scores tends to improve the reliability and tends of improve its validity up to a point. However giving too much weight to more reliable subsections can decrease validity.

Other approaches to creating a composite score include different calibration methods of the subsections. One approach is to calibrate the subtests together using two- or three- parameter uni-dimensional item response models (IRT) where weighting occurs naturally as part of the model. Items better discriminate examines will contribute more information to examinee’s score estimates and have more weights. The main problem of this approach is that if different subtests measure different latent traits then conducing IRT scaling violates the assumption of uni-dimensionality.

Another approach is to use a compensatory system of scoring. The compensatory system is often used when the skills measured by different tests are considered to have some overlap but different skills are allowed to compensate for each other. For example, a compensatory system that combines listening and speaking skills is desirable since both skills are needed in order to communicate effectively. The most common practice for creating compensatory composite scores is to calibrate test items for each two subtests separately using item response theory (IRT) and then create a composite score as a weighted sum or average of two subtests. The composite score is a function of the scores assigned to the two subtests, the relative weights assigned to each subtest, and of the correlations between the subtests. Because of the differential weighting, the exact dimensional definition of the resultant measure might be different from those based on the individual subtest. For example, if one of the subtests receives disproportionably more weight then the trait it measures would be considered to be the predominant trait measured by the composite.

When combining scores from different subtests, the resulting score should be considered as multidimensional. There are many psychometric issues that need to be address when combining scores from different subtests and then attempting to interpret the resulting score from a set of multidimensional subtests models (Reckase & McKinley, 1991; Ackerman, 1994.) First, it is important to know what composite of traits is being measured. Second it is necessary to ensure that all examinees, no matter where they are located in the latent trait space, are being measured on the same composite of traits. Thus, the role of multidimensionality in the interpretation of meaning given to various score levels must be examined.

To further characterize the nature and performance of these methods, this study examines four methods of parameterizing composites, two of which are conducted within the one-dimensional item response modeling framework and two of which are multidimensional approaches. The reliability and fit of these estimates is compared across approaches.

DATA

The data was taken from a State Assessment of English Language Proficiency. The assessment system LAS Links (CTB/McGraw-Hill, 2006) is a contract with a private agency to fulfill requirements under No Child Left Behind. The assessment is given in grades K-12 with five levels of tests designed for K-1, 2nd-3rd grade, 4th-5th grade, 6th-8th grade, and 9th-12th grade students. The students used as a sample for this analysis all completed either Level 2 or Level 3 tests corresponding to their grade being in lower elementary school (2nd-3rd grade) or upper elementary school (4th-5th grade). The assessment contains one hundred items. The content areas covered include Speaking, Listening, Reading, and Writing. The number of items on each section of the test is shown in Table 1.1 below. Only the Speaking and Listening subtests are used for the discussion of composite scores. Further study of composite scoring strategies might include analysis of the four subtests to create and Overall composite score. However, since only the Speaking and Listening subtests are used for this analysis, discussion of the other subtests is discontinued at this point for the rest of the analysis.

Table 1.1: Assessment Content by Subtest

Subtest	Items	Number of Items
Speaking	1 - 20	20
Listening	21 - 40	20
Reading	41-75	35
Writing	76-100	25

The Speaking subtest is designed to measure vocabulary, social and academic language, and more sophisticated grammatically correct verbal expressions as appropriate for the level. The Speaking subtest covers four speaking proficiencies including speaking in words, speaking in sentences, making conversation and telling stories. The speaking subtest individually administered by a fluent English speaker. For lower and upper elementary school students, the examiner reads questions from the student booklet and points to illustrations in a cue picture book. The Speaking subtest thus also requires listening ability. The responses may be single words, phrases, or multiple sentences as required. For example, an item from the "speak in words" group consists of the examiner pointing to an illustration of a crayon and asking, "What is this and what is it used for?" The student would then respond in words or phrases.

The Speaking subtest takes about 10 minutes to administer and is graded in real time by the examiner using a rubric. The answer to the example question above would be scored as "Correct", "Incorrect", or "No Response." Items requiring more complex answers are scored polytomously. Details on the number of levels for each item are shown in Table 1.2. The scoring rubric has levels corresponding to non-response or non-English response, on-topic English words that do not answer the questions, a correct response with grammatical errors, and a correct response with few or no grammatical errors. The levels represent levels of English speaking proficiency. Using this rubric, a hypothesized construct map can be drawn for the Speaking domain. It is shown in relation to the other English proficiency constructs in Figure 1.1.

Figure 1.1

Level	Description	Example Item and Response to Item with an illustration showing a girl, a woman, and a dog.
0	Does not address the topic. Does not respond in English	Tell a Story: El perro y su mama…
1	On-topic English words but does not satisfy the task	Tell a Story: His mom and him, the dog. Him and his mom and the dog…the girl and dog.
2	Adresses the prompot but insufficient and incorrect vocabulary makes overall communication unclear.	Tell a Story: His mom and himm…look the dog. His mom ad his say, "The dog." Him and his mom put the dog in the, the chair. Him and him mom…the girl and dog.
3	Addresses the prompt with overall clear communication despite errors in grammar and vocabulary limitations that create some confusion.	Tell a Story: The boy and his mama see the picture of the dog. They see the dog. She calls. The girl, she, she gets the dog. She is happy.
4	Adresses the prompt woth overall clear communication. Idease and content expressed with ease approaching a native speaker.	Tell a Story: They see a picture of a lost dog. Then they see the dog. The boy, he points to the dog. The mom calls. The girl comes. She gets tne dog and takes him

The Listening subtest is designed to measure general comprehension and inferential and critical thinking skills at a discourse level that integrates academic language as appropriate for the level (CTB/McGraw-Hill, 2006). The listening subtest covers three listening proficiencies including listening for information, listening in the classroom, and listening with comprehension. The listening subtest is administered to a group of students by a fluent English speaker who reads from the test directions or uses the audio CD or cassette. The Listening test takes approximately 15 minutes per group to administer. An example item might show two illustrations; A and B. Illustration A is of a student drawing the ear on a dog. Illustration B is of a student writing the name of the dog and then drawing spots on the dog. The examiner would say, “Listen to my directions. Draw the dog’s ear. Now look at the pictures and circle the answer that shows ‘Draw the dog’s ear,’” (CTB/McGraw-Hill, 2006). Responses are marked in the student response booklet and scored dichotomously. Details on the number of scoring levels for both the Speaking and Listening sections are shown in Table 1.2.

Table 1.2: Number of Scoring Categories for Analyzed Items

Subtest	Item(s)	Scoring	Scores	Number of Score Categories
Speaking	1 - 10	Dichotomous	0, 1	2
Speaking	11-19	Polytomous	0, 1, 2, 3	4
Speaking	20	Polytomous	0, 1, 2, 3, 4	5
Listening	21 - 40	Dichotomous	0, 1	2

Responses from 12,008 lower and upper elementary school students from the Speaking and Listening subtests are used for this analysis. All of the students who participate in this assessment have been identified because of their need to demonstrate English Proficiency. While many students do not complete all of the items, only students with a full set of responses are included in this analysis. This resulted in analysis of 71.6% of the participating lower elementary school students and 84.55% of the participating upper elementary school students. While exploration of the missingness process and its implications for bias in parameter estimates is beyond the scope of this study, it may be informative because a significant number of students were removed from the sample due to missing responses. Details on the sample including gender and ethnicity are shown in Figure 1.3a. The frequencies, row percentages, and column percentages are shown for each combination of gender and ethnicity in the sample and in total. For instance, 47.05% of the students are female while 52.46% are male. Additionally, the largest ethnic group comprises 89.09% of the sample with smaller percentages being split between the other ethnic groups. This information could be useful for further exploration of Differential Item Function DIF or bias in composite scoring schemes but is also beyond the scope of this study. The grade and test level frequencies are shown in Figure 1.3b. It should be remembered that the lower elementary school and upper elementary school levels consist of two grades. This means that if a student in grade 2 fails the English Proficiency examination, he would take a similar examination in grade 3.

Figure 1.3a: Sample Demographics

	Gender			
Ethnicity	**No Response**	**Female**	**Male**	**Total**
1	**0**	**12**	**18**	**30**
	0	*40*	*60*	*100*
	0	0.21	0.29	0.25
2	**0**	**31**	**29**	**60**
	0	*51.67*	*48.33*	*100*
	0	0.55	0.46	0.5
3	**0**	**305**	**414**	**719**
	0	*42.42*	*57.58*	*100*
	0	5.4	6.57	5.99
4	**0**	**10**	**12**	**22**
	0	*45.45*	*54.55*	*100*
	0	0.18	0.19	0.18
5	**5**	**5089**	**5604**	**10698**
	0.05	*47.57*	*52.38*	*100*
	8.47	90.07	88.97	89.09
6	**0**	**101**	**122**	**223**
	0	*45.29*	*54.71*	*100*
	0	1.79	1.94	1.86
7	**0**	**1**	**2**	**3**
	0	*33.33*	*66.67*	*100*
	0	0.02	0.03	0.02
8	**0**	**2**	**3**	**5**
	0	*40*	*60*	*100*
	0	0.04	0.05	0.04
No Response	**54**	**99**	**95**	**248**
	21.77	*39.92*	*38.31*	*100*
	91.53	1.75	1.51	2.07
Total	**59**	**5650**	**6299**	**12008**
	0.49	*47.05*	*52.46*	*100*
	100	100	100	100

Frequencies are shown in bold, row percentages in Italics with column percentages below them.

1.3b: Level and Grade Frequencies

Grade	Lower Elementary (Level 2)	Upper Elementary (Level 3)	Total
2	3,212		3,212
3	3,532		3,532
4		2,697	2,697
5		2,567	2,567
Total	**6,744**	**5,264**	**12,008**

In addition to item responses and demographic information, each student has an estimated scale score based on previous calibrations of the data. Scale scores correspond to proficiency levels in each domain which are used for decision making purposes. In addition, in previous analysis, composite scale scores were derived using averages of the scores from the individual domains. While the scoring strategies used in the previous analyses are not employed for this analysis, examination of the relationship between the scores gives useful information as to the possible nature of the relationship between the domains.

Original IRT Calibration

The uni-dimensional IRT techniques were used to calibrate, scale, and place the LAS Links items onto the LAS Links scale to assure comparability of scores from form to form. Since both multiple-choice (MC) and constructed-response (CR) items are included in the test, both item types were calibrated together and placed on a single scale, using the three-parameter logistic (3PL) model (Lord & Novick, 1968) and the two-parameter partial-credit model (2PPC) (Muraki, 1992; Yen, 1993). The 3PL model was used for the multiple-choice items because it estimates student guessing in addition to item location (difficulty) and allows for differences in item discrimination. The parameters were estimated simultaneously for dichotomous and polytomous items using marginal maximum-likelihood procedures implemented via the Expectation Maximization (EM) Algorithm.

Listening and Speaking items were calibrated separately. In addition, an Oral composite scale consists of both Listening and Speaking items was created by calibrating both subtests together to create an oral composite scale. The Pearson correlations between the scale scores for each level are shown in Table 1.4 below. Across the two levels it is seen that similar patterns in terms of the correlation between the scale scores exists. For instance, for both levels, the Speaking scale score is highly correlated with the Oral scale score with values of .91 and .92 respectively. One may begin to think that the nature of the latent ability is responsible for this high correlation. However, it is important to note that the Speaking test has a larger number of possible score points than the listening test because it is polytomously scored. Since the Oral scale score in this case was calculated by a simultaneous calibration of all Speaking and Listening items, it is not surprising to find that Speaking test dominates the Oral scale score. This result gives further motivation to explore multidimensional scoring strategies versus simple averages. It should also be noted that the correlation between the Listening and Speaking is only moderate for both levels with values of .42 and .45. This gives evidence that some information about the two domain abilities may be contained in the results of each subtest, further motivating a multidimensional approach.

Table 1.4: Correlations between Scale Scores

Level		Speaking	Listening	Oral	Overall
Lower Elementary School	**Speaking**	1			
	Listening	0.42	1		
	Oral	0.91	0.69	1	
	Overall	0.72	0.70	0.79	1
		Speaking	**Listening**	**Oral**	**Overall**
Upper Elementary School	**Speaking**	1			
	Listening	0.45	1		
	Oral	0.92	0.72	1	
	Overall	0.74	0.79	0.86	1

METHOD

Exploratory Factor Analysis

While correlations between the previously calculated scale scores are useful for Preliminary exploration, a deeper correlational analysis can further exploration of the multidimensional nature of the domains. More directly, a factor analysis of the correlation matrices can help to determine whether the composite domains are truly multidimensional, and if so, the number of dimensions they might comprise. Factor analysis of correlation matrices versus the actual data allows for factor analyses which take into account the fact that item response data is not continuous. The polychoric and its special case for dichotomous scores, the tetrachoric correlation have been used to estimate correlations between ordered category data, such as item responses scored in levels (Pearson, 1901). In measurement the tetrachoric and polychoric correlations are considered to be appropriate tools with the underlying ability is considered continuous but the responses have been divided into ordered categories. In addition certain assumptions must be met.

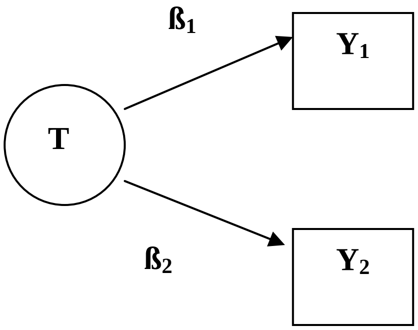
To further elaborate the model and its assumptions as outline in the terms of Classical Test theory, the model and assumptions of tetrachoric correlation are outlined. The dichotomous case is used for simplicity of presentation. Consider two item responses $\mathbf{X_1}$ and $\mathbf{X_2}$ which are the scored responses to two different items. Further say that $\mathbf{Y_1}$ and $\mathbf{Y_2}$ be the continuous latent ability values underlying the responses $\mathbf{X_1}$ and $\mathbf{X_2}$for each item. In other words, $\mathbf{X_1}$ and $\mathbf{X_2}$ are discretized versions of $\mathbf{Y_1}$ and $\mathbf{Y_2}$. Finally consider the true latent trait level to be **T**. Thus the measurement expression for each item can be written such that the observed levels $\mathbf{Y_1}$ and $\mathbf{Y_2}$are products of the true level and a coefficient plus some unique variance and random error, included in $\mathbf{e_1}$ and $\mathbf{e_2}$. This can be written as shown below.

$$\mathbf{Y_1 = ß_1T + e_1,}$$
$$\mathbf{Y_2 = ß_2T + e_2.}$$

A few key assumptions for use of the tetrachoric correlation are then that the distribution of **T** is normal. Also it is assumed that the error terms are also normally distributed and that they are independent across items. It follows that $\mathbf{Y_1}$ and $\mathbf{Y_2}$are also normally distributed. For simplicity it is also often assumed that variance of the error terms is equal and that all assumed normal distributions are standard normal distributions. If the assumptions have been sufficiently met, it can be said that the tetrachoric correlation, **r**, is the product of the coefficients.

$$\mathbf{r = ß_1ß_2}$$

The idea behind this example can be further illustrated using path diagrams. Following the rules of path analysis, the correlation between $\mathbf{Y_1}$ and $\mathbf{Y_2}$ is the product of the path coefficients or $\mathbf{ß_1ß_2}$.



There is an obvious advantage in interpretive ability in using this version of the correlation as shown as an example in the path diagram. Some additional advantages of using polychoric correlations are that the correlation does not depend on the number of levels and they are easily estimable with available software. The fact that the number of categories does not challenge the validity of the estimated correlation is especially helpful in the case where items with different numbers of categories are analyzed, such as is the case with the sample used in this analysis.

Once the correlation matrices have been estimated, a factor analysis of this correlation can be conducted with an additional input of the original sample size in many statistical software packages including SAS, Stata, and PRELIS often in the same package where the correlation matrices were estimated. Factor analysis using the principal axis factoring method is chosen for the example in this analysis. Since the focus of this paper is not factor analysis, but an exploration of multidimensional IRT models, simply using factor analysis as a preliminary step, the full factor analytic models are not presented here. It is important however to note some key factor analytic tools that can be used for the purposes of this analysis. First, to answer questions as to the number of appropriate factors can be answered using the generally accepted techniques such as the Scree Plot, or a plot of the eignenvalues, examination of the eignevalues and AIC and BIC indices. To answer questions as to the nature of the dimensionality, the factor loadings can be examined by looking at their values and by noting the amount of variance and uniqueness of each factor. With the help of evidence from the factor analysis results, multidimensional models can be examined.

Multidimensional Item Response Model

There are a total of four calibration models for calibrating oral items were examined and compared for model-data fit for each population: one uni-dimensional model (UIRT) and three two-dimensional models (MIRT). The BMIRT (Yao, 2004), a computer program that is capable of calibrating MIRT models, was used to estimate parameters both UIRT and MIRT models. BMIRT was used to recalibrate the oral composite scale in order to obtain model-data fit statistics that would be used in model comparison. Yao's BMIRT estimation program has been studied through both simulation and real data applications (Yao 2004, Yao & Schwartz 2005.) BMIRT uses a Baysian formulation of multivariate item response theory. This program uses Markov chain Monte Carlo (MCMC) method to estimate the item, ability, and latent parameters for multidimensional multi-group methods for both dichotomous and polytomous data in either exploratory or confirmatory modes. The parameter recovery studies of BMIRT can be found in Patz and Yao (2003) and Yao and Boughton (2005.)

Estimation can be problematic with multidimensional IRT models due to the number of parameters needed to be estimated in the model compared with uni-dimensional IRT model. MCMC estimation method avoids taking the derivative from the posterior distributions which can be problematic when a large number of dimensions are present in the data. The item, ability, latent parameters were estimated using Metropolis-Hastings algorithm. Metropolis-Hastings is a general algorithm commonly used for Markov chain simulation methods which involves drawing samples from appropriate distributions and then correcting these draws to better target the posterior distribution. Technical details regarding the estimation procedures used in BMIRT for mixed-item format test can be found in Yao and Schwartz (2005.)

Compensatory multidimensional model was used to estimate the constructed response items presented as a two-parameter partial credit model and estimate the multiple-choice items as a three-parameter item response model. Under the compensatory model, having a higher ability on one dimension can potentially compensate for lower ability on a second dimension (Reckase 1985.)

As with factor analysis, there are two basic approaches to MIRT analysis- exploratory and confirmatory. In exploratory procedures, the emphasis is on discovering the best fitting model, while in confirmatory approaches the focus is on evaluating the extent to which the data follow a hypothesized model developed a priori on the basis of content and process analysis of the instrument to be analyzed. The exploratory procedure was used in this research.

The proposed models varied both in number of dimensions and factor correlations. Models with one and two dimensions were estimated. The one dimensional model assumes that Listening and Speaking items measure one single latent dimension.

The two dimensional model assumes that Listening items measure one dimension while the Speaking items measure a different dimension. Although it is possible to impose a simple structure for this analysis, it was determined that a simple structure might be too restrictive. Therefore, all items are free to load on either dimension. For the two-dimensional case, three different factor correlations were used, r=0.0, 0.3, and 0.5. The choice of correlations is arbitrary but it is also used to fix the scale of the latent dimensions.

Like the one-dimensional IRT models, MIRT models are under-identified. The metric or scale needs to be fixed in the estimation process to solved this indeterminacy problem. Following the recommendations from Yao and Schwarz (2005), the population parameters have been fixed as a normal distribution with mean of 0 and standard deviation of 1 for the uni-dimensional case. For the two-dimensional models, the population parameters have been fixed as a multinormal with mean (0,0), and the variance and covariance matrix as,

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

where r=0.0, 0.3, or 0.5.

For each model, 10,000 iterations were specified, with the first 500 of them assigned to the “burn-in” phase with arbitrary starting points. BMIRT program evaluates the priors and proposal functions used in the run in terms of their reasonable acceptance rates in the MCMC sampling. The acceptance rates for all the runs are between 30 and 39 percents.

Model Evaluation

Solutions for the four calibration models for both Lower and Upper Elementary populations were obtained. The best fitting multidimensional IRT model would be chosen to create multidimensional IRT scores for students in both populations.

The model-data fit of completing models of test structure were compared based on the work of Akaike (1973). The approach is based on a criterion called the entropic information criterion, also know as the AIC, and it involves evaluating model fit in terms of the natural algorithm of the likelihood of the model. The greater the likelihood, the closer the fitted model is presumed to approximate the true model. This approach can be used in situation where various completing models are not nested and therefore can not be compared using the chi-square procedure. The AIC statistics is given by

$$AIC=-2\ log(L)+\ 2k$$

Where the *log (L)* denotes the natural log of the likelihood and k is the number of parameters estimated. The *2k* term constitute a penalty function that penalizes over parameterizations.

The model-data fit was also evaluated using chi-square difference test. Solutions for simple models, such as the uni-demensional model, are obtained first. Models with increasing complicity are then created by adding parameters. These more complex models subsume simpler models, making it possible to test the significance of the contribution of the additional parameters using a chi-square procedure. While it is doubtful that the likelihood ratio statistics produces by programs such as BMIRT is actually distributed as a chi-square, the difference between the values of likelihood ratio chi-square statistics for subsuming models has been shown to be asymptotically distributed as chi-square (Haberman, 1977). The degrees of freedom for the differences between two values of likelihood ratio statistics are equal to the difference between the two models. For the IRT models, this equals to the difference in the number of parameters estimated (Yao & Schwartz, 2005.)

Comparisons of Test Information

Information can be thought of as the inverse of the standard error of approximation. Thus when the information is greater, the approximations are less erroneous, or better. Often the range of maximum information is the area in which cut-points are set for high stakes exams for this reason. The one-dimensional and multidimensional information functions are examined in order to further compare the models.

The one-dimensional formula for the information functions can be extended to the multidimensional case. The multidimensional extensions of the one-dimensional information functions for the 2PPC and 3PL Model are presented here. Using the multidimensional 2PPC Model, the probability, P_{ijk} ,that examinee *i* receives a score of *k-1* on item *j* with K_j possible scores, is given as shown.

$$P_{ijk} = P(x_{ij} = k-1 \mid \vec{\theta}_i, \vec{\beta}_j) = \frac{e^{((k-1)\vec{\beta}_{2j} \circ \vec{\theta}_i - \sum_{t=1}^{k} \beta_{\delta_t j})}}{\sum_{k=1}^{K_j} e^{((k-1)\vec{\beta}_{2j} \circ \vec{\theta}_i^T - \sum_{t=1}^{k} \beta_{\delta_t j})}} \qquad (1)$$

where

- $x_{ij} = 0,...,K_j - 1$ is the score received by examinee *i* on item *j*.
- $\vec{\theta}_i = (\theta_{i1},...,\theta_{in})$ is a vector of ability parameters in *n* dimensions.
- $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta_1 j},...,\beta_{\delta_{K_j} j})$ is a vector of item parameters.
- $\vec{\beta}_{2j} = (\beta_{2j1},...,\beta_{2jn})$ is a vector of discrimination parameters in *n* dimensions.
- $\beta_{\delta_1 j},...,\beta_{\delta_{K_j} j}$ are the threshold parameters with $\beta_{\delta_1 j} = 0$.

The item information function can be written as shown.

$$I_j(\vec{\theta}) = \sigma_j^2 (\vec{\beta}_{2j} \times \vec{\beta}_{2j}) \qquad (2)$$

where

$$\sigma_j^2 = \sum_{k=1}^{K_j} (k-1)^2 P_{jk} - \left(\sum_{k=1}^{K_j} (k-1) P_{jk} \right)^2$$

Using the multidimensional 3PL Model, the probability, P_{ij1} ,that examinee *i* answers item *j* correctly is given as shown.

$$P_{ij1} = P(x_{ij} = 1 | \vec{\theta}_i, \vec{a}_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + e^{\left(-D\vec{a}_j \circ \vec{\theta}_i^T + b_j\right)}} \qquad (3)$$

where

- $\vec{\theta}_i = (\theta_{i1},...,\theta_{in})$ is a vector of ability parameters in *n* dimensions
- b_j is a scale difficulty parameter
- $\vec{a}_j = (a_{j1},...,a_{jn})$ is a vector of discrimination parameters in *n* dimensions.
- c_j is a scale guessing parameter

The item information function is written as shown.

$$I_j(\vec{\theta}) = \frac{P_{j1} Q_{j1}}{\left(1 + c_j e^{(-D\vec{a}_j \circ \vec{\theta}^T + b_j)}\right)^2} (\vec{a}_j \times \vec{a}_j) \qquad (4)$$

The item information from the models for constructed response or multiple choice items can be added over items to create the test information. The test information functions have been plotted across ability levels.

Student’s Ability Estimates

After obtaining estimates of item parameters, expected posterior (EAP) estimates of abilities were obtained for both the UIRT and MIRT models. For each student, three different ability estimates for obtaining composite oral scores are available, one based on simple average of Speaking and Listening scale, one based on the concurrent calibration of Speaking and Listening scales, and one based on MIRT scale. Note that in the interest of time, the simple average of Speaking and Listening ability estimates were obtained from the original IRT calibration using a different IRT software than BMIRT.

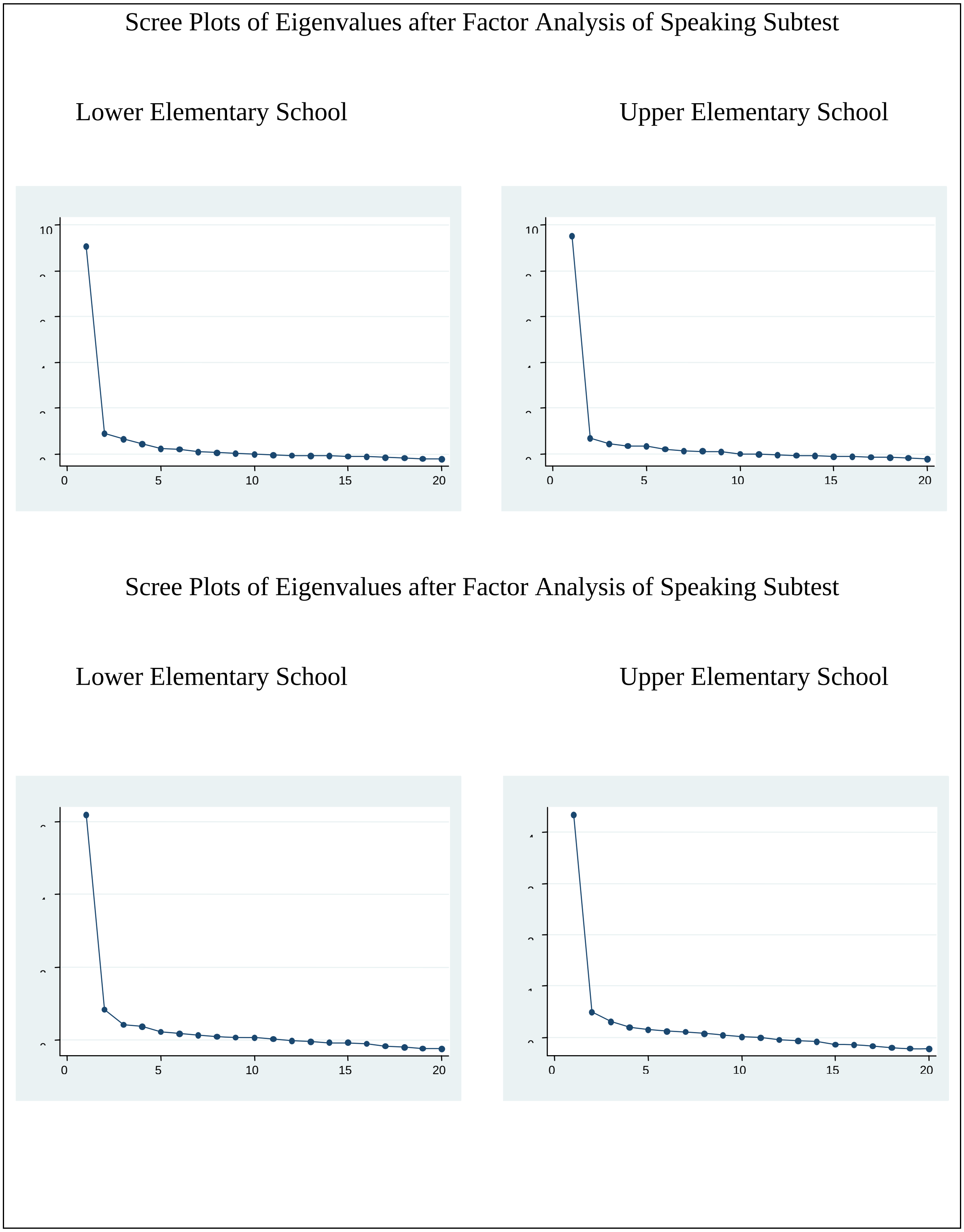
Since these three ability estimates are not on the same scale, direct comparisons among the three estimates are not possible. To examine the relationship of the ability estimates generated by three different ways of creating composite scores, the correlation coefficients among the three different composite scores were computed.

RESULTS

Exploratory Factor Analysis

It is generally assumed for the purposes of calibration with one-dimensional item response models, that each single subtest describes a single latent ability. To asses the dimensionality of each of the two pertinent subsets, a factor analysis of each subtest was conducted in Stata after having computed the polychoric correlation matrices using the polychoric routine (Kolenikov, 2004). The Scree Plots from these analyses are shown below.

Figure 2.1: Scree Plots



Scree Plots of Eigenvalues after Factor Analysis of Speaking Subtest

Lower Elementary School

Upper Elementary School

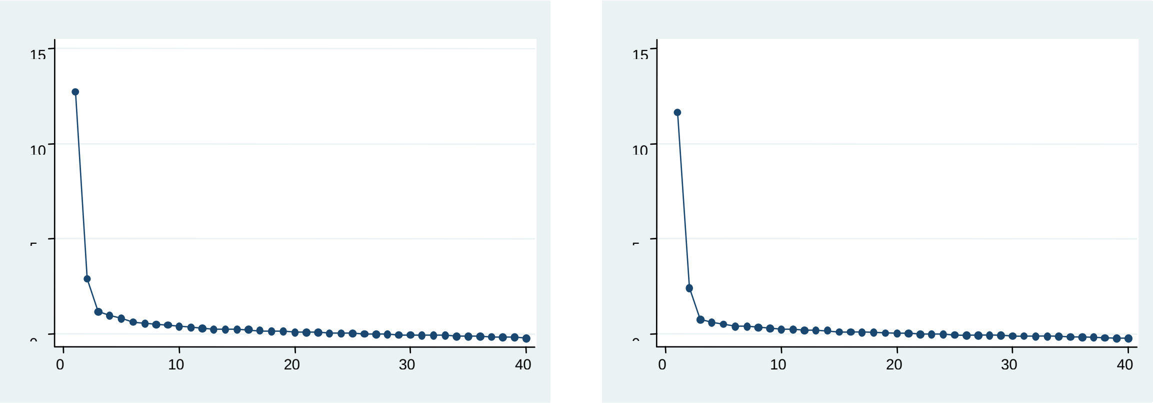
Scree Plots of Eigenvalues after Factor Analysis of Speaking Subtest

Lower Elementary School

Upper Elementary School

Scree Plots of Eigenvalues after Factor Analysis of Composite Domain Lower

Elementary School Upper Elementary School



It is seen through an examination of the elbow of the Scree Plots that for both Lower and Upper Elementary samples, that each single domain is one-dimensional and that the composite domains are possibly two-dimensional. This was further confirmed when examining the eigenvalues. Thus it is determined that the two-factor solution is the best solution for the Composite domain for both Lower and Upper Elementary school students.

Since the number of appropriate factors has been determined, it is not necessary to examine the nature of the relationship of the factors given the two-factor solution. The factor loadings for each item are shown in Figure 2.2. Items with factor loading of 0.4 or higher were considered having a significant factor loading on the factor. For Lower Elementary school population, all the Speaking items had a positive and significant loading on the first factor and a negative but insignificant loading on the second factor.

The majority of the Listening items had positive and significant loadings on the first factor but these items also had positive and moderately high loadings on the second factor. There are a few Listening items that did not have significant high loadings in either factor. Similar patterns were found for the Upper Elementary school population. Lastly, Listening items for the Upper Elementary school seem to exhibit more double loadings on both factors as compared to the Lower Elementary school.

Since the factor extraction method first finds a dominant factor and then secondary factors, the interpretation may not directly correspond to the type of dimensional interpretation used in IRT analyses. Indeed the results show that the first factor is heavily affected by the construct common to all the Speaking items and that the second factor measures a somewhat different construct from the first factor. The second factor seems to more reflective of the construct represented by the Listening items. More over, there might be a negative correlation between the first factor and the second factor. Thus we see the pattern expected from a factor analysis.

The intent of the factor analysis is both to verify that the data is multidimensional and to describe the dimensionality preliminarily before the exploratory analysis using Multidimensional Item Response Theory. Given the results, the two-dimensional structure has been verified. The dimensionality structure has been questioned as a point of interpretational difference between the varying intent of ability estimation. More explicitly, for the purpose of defining separate scores for a speaking and listening dimension, it would be disconcerting to find that the combination of these two sections resulted in one dominant factor and one secondary factor. However, for the purposes of defining and oral composite score, this result seems natural. In fact that dominant factor could be considered to represent the bulk of a student's Oral ability. Since it has been pointed out the production of a dominant and secondary factor is a result of the factor extraction method, the multidimensional analysis should be used to provide more information on the dimensional structure using the IRT framework as is done in the following sections. Furthermore, based on the findings from the exploratory factor analysis where it was found that the Speaking items load negatively on the second dimension, the compensatory MIRT model seems to a reasonable model to use in the multidimensional IRT calibration.

Figure 2.2: Factor Loadings and Uniqueness from Two-Factor Solution.

Factor Loadings
For Lower Elementary School Two-Factor Solution in Composite Domain

Item	Skills	Item Type	Level	Factor 1	Factor 2
1	Speaking	CR	2	**0.563**	-0.1759
2	Speaking	CR	2	**0.5987**	-0.2794
3	Speaking	CR	2	**0.4922**	-0.2203
4	Speaking	CR	2	**0.6578**	-0.2747
5	Speaking	CR	2	**0.582**	-0.1585
6	Speaking	CR	2	**0.6099**	-0.2449
7	Speaking	CR	2	**0.5438**	0.034
8	Speaking	CR	2	**0.7047**	-0.1181
9	Speaking	CR	2	**0.532**	-0.0041
10	Speaking	CR	2	**0.5325**	-0.1929
11	Speaking	CR	4	**0.7476**	-0.2371
12	Speaking	CR	4	**0.8038**	-0.2625
13	Speaking	CR	4	**0.6776**	-0.389
14	Speaking	CR	4	**0.6342**	-0.3042
15	Speaking	CR	4	**0.7086**	-0.2915
16	Speaking	CR	4	**0.6438**	-0.1975
17	Speaking	CR	4	**0.668**	-0.2192
18	Speaking	CR	4	**0.652**	-0.2431
19	Speaking	CR	4	**0.6805**	-0.2644
20	Speaking	CR	5	**0.6545**	-0.2378
21	Listening	MC	2	**0.494**	0.267
22	Listening	MC	2	**0.655**	0.3055
23	Listening	MC	2	**0.6137**	0.3645
24	Listening	MC	2	**0.4985**	**0.4228**
25	Listening	MC	2	**0.4567**	**0.4119**
26	Listening	MC	2	**0.5279**	**0.4346**
27	Listening	MC	2	0.3585	0.1958
28	Listening	MC	2	**0.8243**	0.3041
29	Listening	MC	2	**0.3134**	0.217
30	Listening	MC	2	**0.5457**	0.3876
31	Listening	MC	2	**0.4107**	0.2963
32	Listening	MC	2	**0.4131**	0.2381
33	Listening	MC	2	**0.5519**	0.2988
34	Listening	MC	2	0.3267	0.1966
35	Listening	MC	2	0.3046	0.2311
36	Listening	MC	2	0.2879	0.2006
37	Listening	MC	2	**0.5931**	0.3155
38	Listening	MC	2	0.2747	0.1749
39	Listening	MC	2	0.3129	0.2518
40	Listening	MC	2	0.3742	0.294

Note: Level is the number of score categories for the item.

Factor Loadings
For Upper Elementary School Two-Factor Solution in Composite Domain

Item	Skills	Item Type	Level	Factor 1	Factor 2
1	Speaking	CR	2	**0.7589**	-0.0877
2	Speaking	CR	2	**0.6642**	-0.2199
3	Speaking	CR	2	**0.5791**	-0.0142
4	Speaking	CR	2	**0.6812**	-0.1658
5	Speaking	CR	2	**0.5137**	-0.0177
6	Speaking	CR	2	**0.6327**	-0.2072
7	Speaking	CR	2	**0.8085**	-0.083
8	Speaking	CR	2	**0.5333**	-0.2035
9	Speaking	CR	2	**0.706**	0.0207
10	Speaking	CR	2	**0.6747**	-0.212
11	Speaking	CR	4	**0.6313**	-0.3163
12	Speaking	CR	4	**0.6709**	-0.2983
13	Speaking	CR	4	**0.7531**	-0.1811
14	Speaking	CR	4	**0.6997**	-0.1345
15	Speaking	CR	4	**0.6757**	-0.1601
16	Speaking	CR	4	**0.6332**	-0.197
17	Speaking	CR	4	**0.721**	-0.1599
18	Speaking	CR	4	**0.7076**	-0.1904
19	Speaking	CR	4	**0.6222**	-0.1656
20	Speaking	CR	5	**0.6821**	-0.1914
21	Listening	MC	2	0.1944	0.1898
22	Listening	MC	2	0.115	0.1744
23	Listening	MC	2	0.2397	0.2494
24	Listening	MC	2	**0.4624**	0.3182
25	Listening	MC	2	**0.456**	0.3783
26	Listening	MC	2	**0.4986**	**0.4037**
27	Listening	MC	2	**0.534**	**0.4386**
28	Listening	MC	2	**0.4466**	0.3727
29	Listening	MC	2	0.223	0.2053
30	Listening	MC	2	**0.4917**	0.3767
31	Listening	MC	2	0.2643	0.2343
32	Listening	MC	2	**0.4263**	0.3155
33	Listening	MC	2	0.2816	0.2036
34	Listening	MC	2	**0.5312**	0.3578
35	Listening	MC	2	0.2988	0.3194
36	Listening	MC	2	0.2929	0.2623
37	Listening	MC	2	0.2808	0.2206
38	Listening	MC	2	0.2031	0.1793
39	Listening	MC	2	0.2603	0.2523
40	Listening	MC	2	0.3406	0.2214

Summary of Model Fit for UIRT and MIRT Calibrations

Table 2.1 summarizes the model selection criterion and results. For each population, Table 2.1 shows the AIC and likelihood ratio chi-square for each of the hypothesized test structures. The AIC with the lowest value could be designated as the preferred model. For Lower Elementary population, the one-dimensional model provided the best fit to the model according to the AIC criterion. While for Upper Elementary population, the two-dimensional model with the correlation between the latent dimensions set to 0.0 provided the best fit to the model, which was better than the two-dimensional model with correlation between the latent dimensions set to 0.3 and 0.5. The chi-square differences were future examined, contrasting one dimension with the three two-dimensional models. For both populations, the chi-square differences are largest between the one dimension and two dimension models with correlation between the latent variables fixed at 0.0.

For the Lower Elementary population, the results of the model-data fit based on the AIC and chi-square difference are not consistent. The inconsistencies in the two model-fit statistics are interesting and deserve further investigations. For the Upper Elementary population, however, the AIC and the chi-square difference tests produce consistent results. Since the main purpose of this research is to investigate how to use the multidimensional model to produce a composite score for the oral scales, the two-dimensional model with the correlation between the latent dimensions fixed to 0.0 was used for both populations to create MIRT composite scores. The possible ramification and implications of the observed inconsistencies in the two model-fit statistics for the Lower Elementary populations would be further discussed in the discussion section.

Table 3. 1. Model Fit for the Models Examined for Oral Scale

Model	Akaike Information	Log-Likelihood	$\chi 2$	df	Number of Item Parameters
Lower Elementary					
I. One Dimension	201,348	-93,809	187,616	6,865	121
II. Two Dimensions (r=0.0)	204,634	-88,668	177,828	13,649	161
III. Two Dimensions (r=0.3)	205,106	-88,903	177,336	13,649	161
IV. Two Dimensions (r=0.5)	206,830	-89,766	179,532	13,649	161
I-II			10,298		20
I-III			9,801		20
I-IV			8,084		20
Upper Elementary					
I. One Dimension	227,936	-108,583	207,170	5,383	121
II. Two Dimensions (r=0.0)	227,550	-103,086	206,172	10,689	161
III. Two Dimensions (r=0.3)	228,342	-103,482	206,964	10,689	161
IV. Two Dimensions (r=0.5)	228,968	-103,795	207,590	10,689	161
I-II			10,998		20
I-III			10,206		20
I-IV			9,580		20

Comparison of Test Information

The results in terms of test information from the one-dimensional model for oral scale are shown in Figure 3.1 (LES) and Figure 3.2 (UES). Given that the model was one-dimensional the ability, Theta, can be interpreted as a combination of Speaking and

Listening abilities. In both cases, the peak amount of information is shown in the middle of the ability range.

The results from the best fitting two-dimensional BMIRT model are shown in two ways. First, the test information using equations (2) and (4) was projected onto the Speaking and Listening dimensions. Since the information is based on the two-dimensional model, the projection plots are rough. For the purposes of presentation, the plots were smoothed using the LOWESS method with a bandwidth of 0.8. In addition to the projection plots, two-dimensional surface plots of the test information are presented for both levels. These plots show the amount of information on the vertical axis across the values of Speaking ability and Listening ability dimensions.

Figure 3.1.

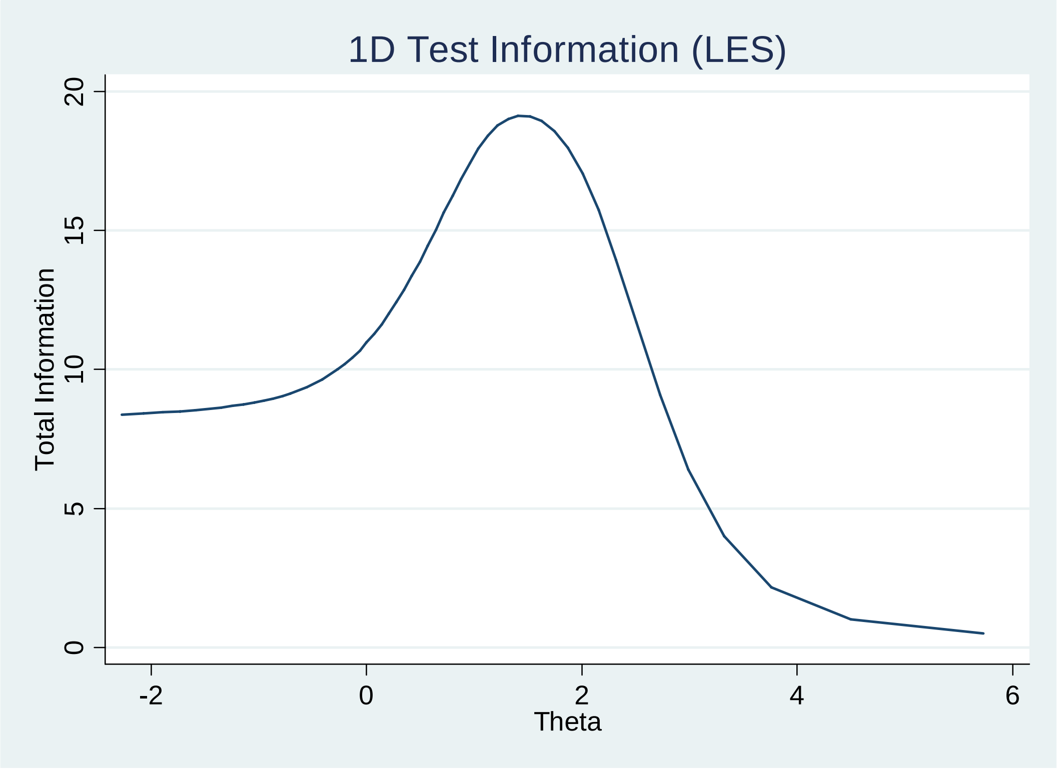


Figure 3.2

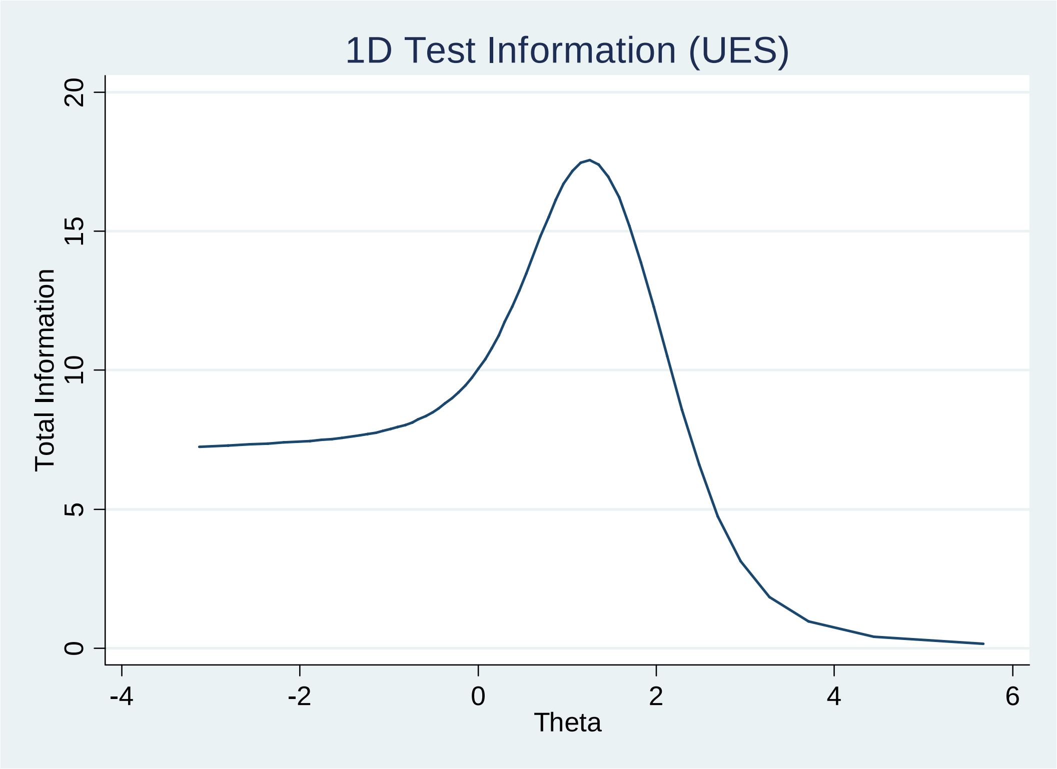


Figure 3.3 shows the two-dimensional test information from the Lower Elementary School calibration onto the Speaking ability dimension while Figure 3.4 shows the same as projected onto the Listening ability dimension. Figures 3.5 and 3.6 shown the same plots for the Upper Elementary School calibration. Since the exploratory multidimensional calibration resulted in many speaking items and listening items showing discrimination in both dimensions, it should be observed that the test information will show amounts of information when projected onto either dimension. If the multidimensional nature of the items had been ignored, plots onto the dimensions would look the same as the one-dimensional plots. Thus the multidimensional calibration is shown in these plots to provide additional information. Also it should be noted that the information projections onto the speaking dimension show greater peaks and thus more

information. This could be as a result of the fact that speaking items were administered in the information rich constructed response format versus the multiple choice format in which the listening items were administered.

Figure 3.3.

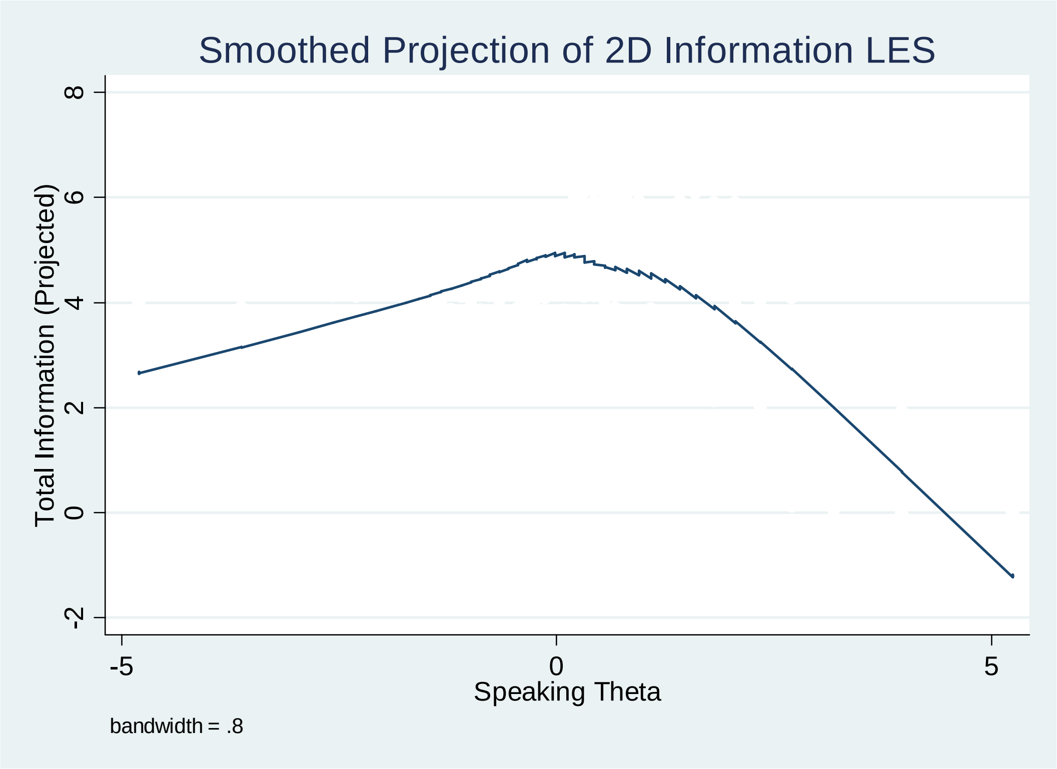


Figure 3.4.

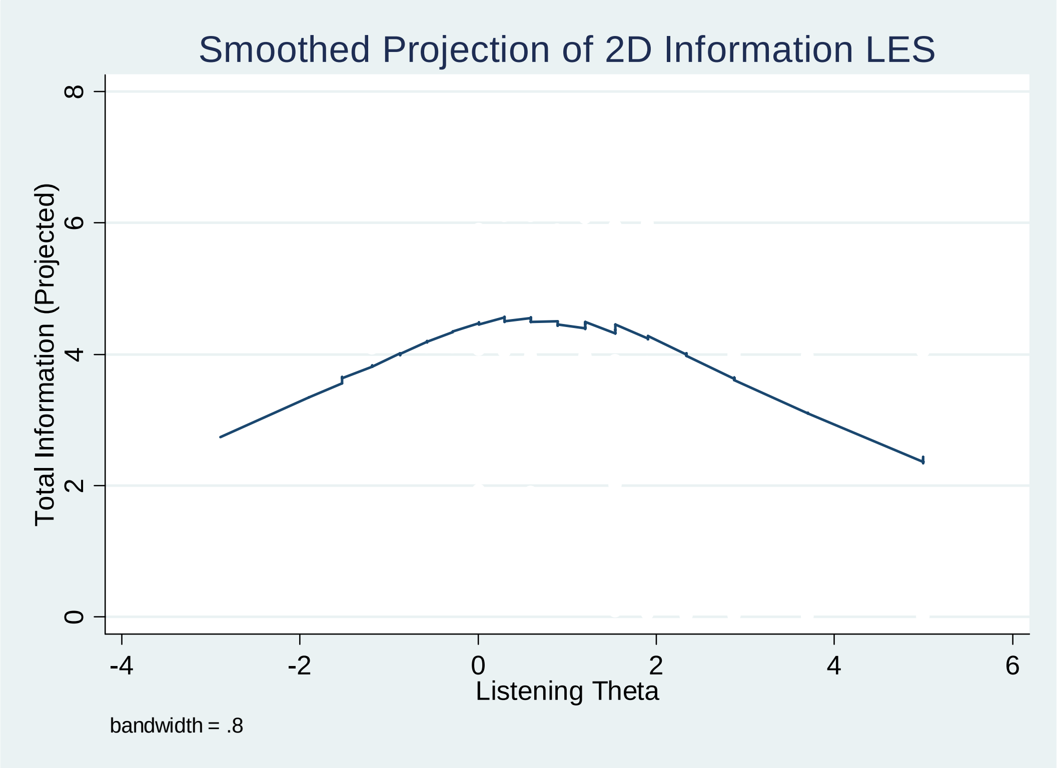


Figure 3.5

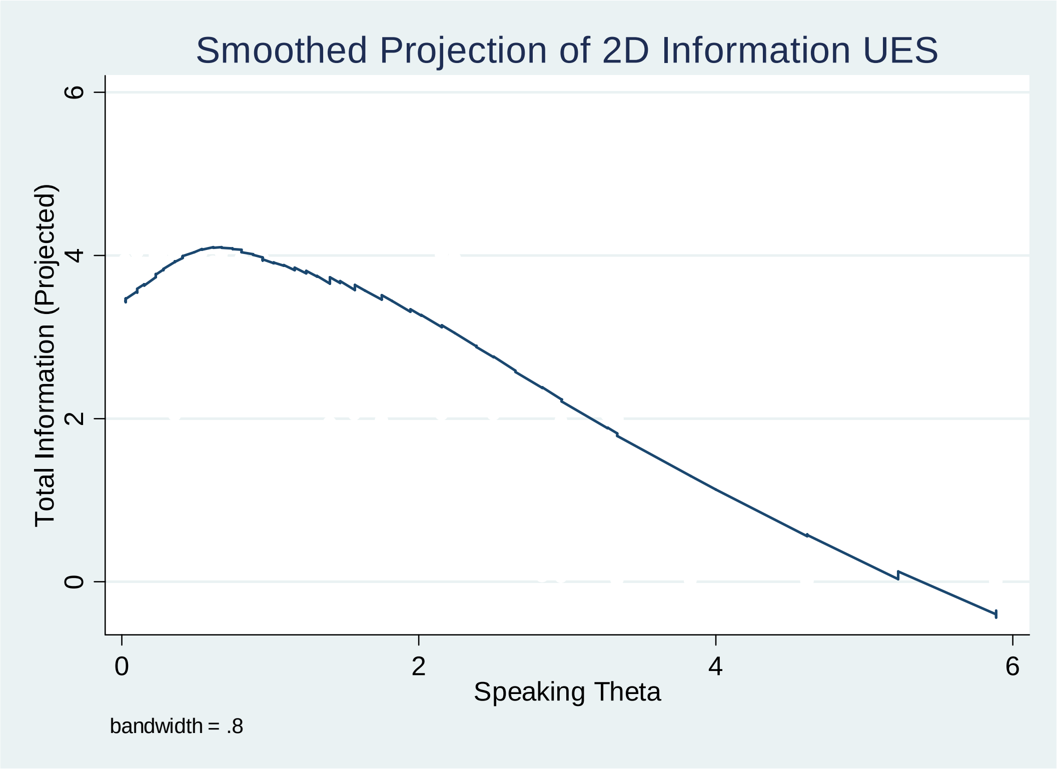
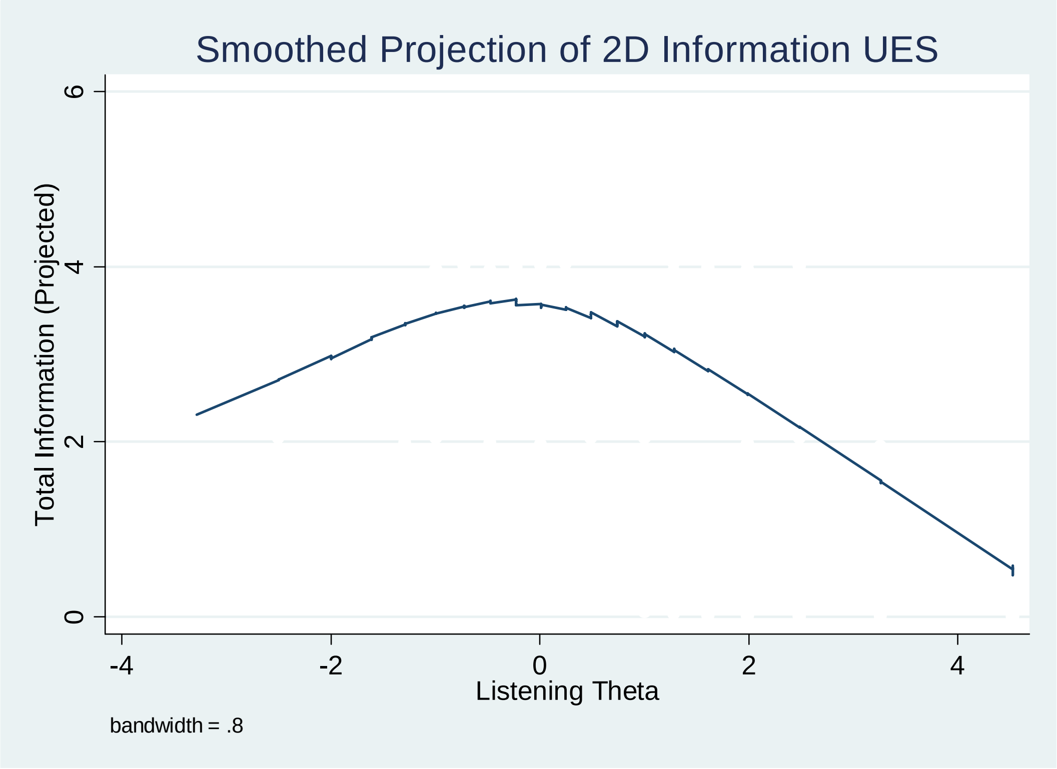


Figure 3.6.



The multidimensional surface plots for the Lower Elementary School and Upper Elementary School calibrations are shown in Figures 3.7 and 3.8. The plots were produced using sample students with varying combinations of speaking and listening abilities based on the BMIRT calibration. Again, the information function shape is seen with the greatest amount of information lying in the middle ranges of ability combinations. Various rotations of the surface plot could be presented to further illuminate the nature of information given by the two subtests. A rotation for instance, could even out some of the peaks shown for the Lower Elementary calibration. However, that analysis is beyond the scope of this study.

Figure 3.7.

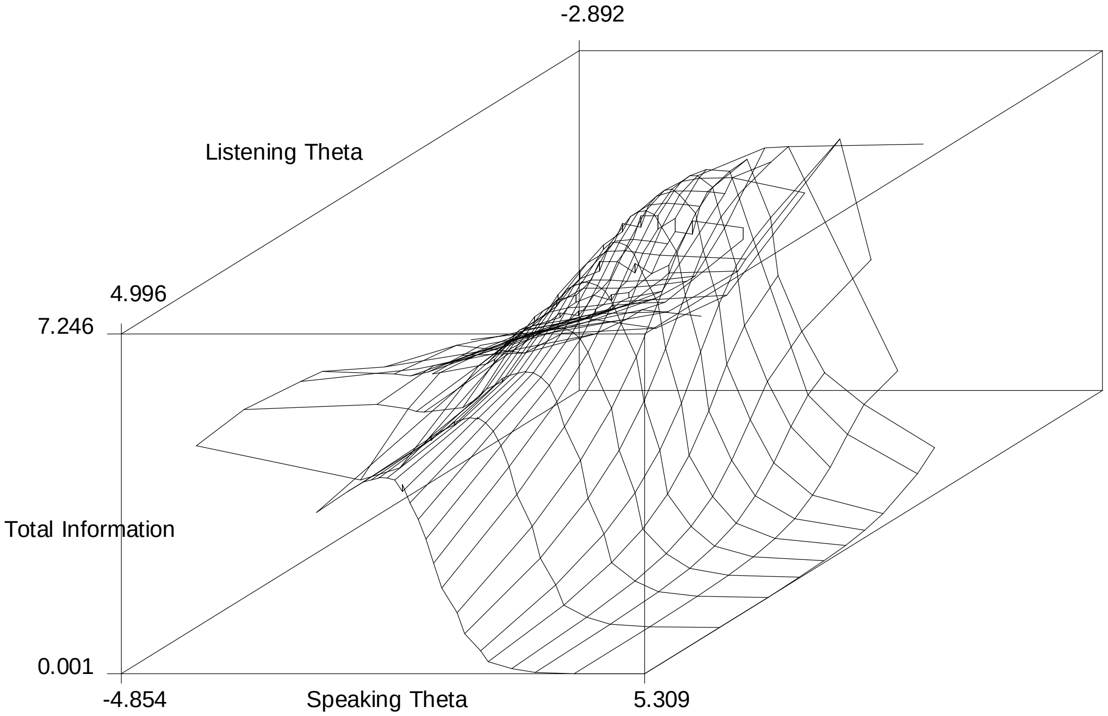
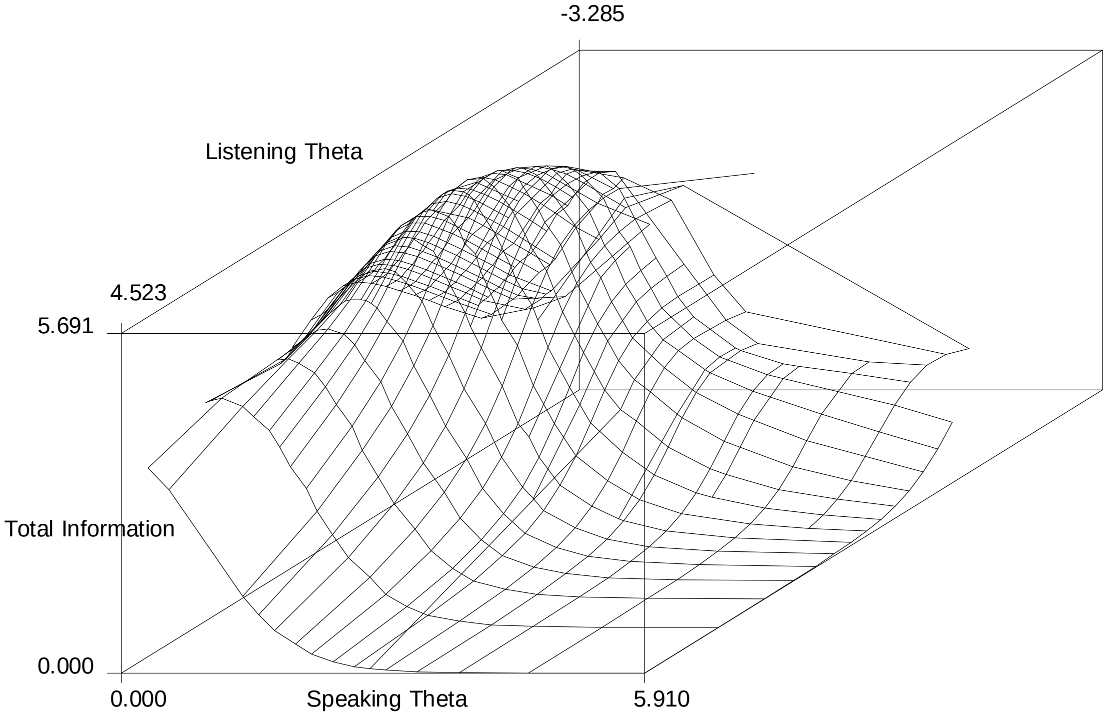


Figure 3.8.



Ability Estimates

After obtaining estimates of item parameters, expected posterior (EAP) estimates of abilities were obtained. For each student, three types of composite scores are available, one based on simple average of Speaking and Listening UIRT estimates, one based on the concurrent calibration of Speaking and Listening items (oral composite estimates), and one based on MIRT estimates. Since these three ability estimates are not on the same scale, direct comparisons among the three estimates are not possible. To examine the relationship of the three composite scores, the correlation coefficients of the three types of composite theta estimates were computed. For ease of presentation, the correlation between the Oral theta estimates and two other composite theta estimates were examined. Furthermore, the individual theta estimates based on separate Speaking and Listening UIRT calibration were also presented in Table 3.2.

Table 3.2 indicates that for the Lower Elementary population, the correlation between the first MIRT theta estimates and the UIRT composite (.77) was slightly lower than those between the second MIRT theta estimates and the UIRT composite (.87.) Furthermore, the size of that correlation is similar to those between the UIRT average and the UIRT composite (.78.) This might be an indication that the average UIRT estimates was highly affected by the first latent dimension of the MBIRT model. It is also interesting to find that the UIRT oral composite correlates significantly higher with the UIRT Speaking theta estimates as opposed to the UIRT Listening theta estimates. Similar patterns were observed for the Upper Elementary population.

Table 3. 2. Correlation Coefficient between the Oral Composite and Five Other Ability Estimates

Population	MBIRT ϑ_1	BMIRT θ_2	UIRT Average	UIRT Speaking	UIRT Listening
Lower Elem.	.77	.87	.78	.82	.25
Upper Elem.	.86	.81	.83	.87	.35

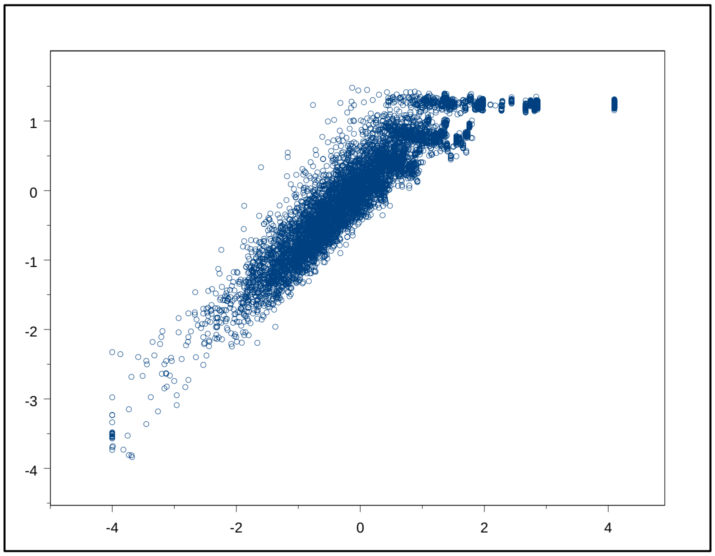
Figure 3.9 and Figure 3.10 present the point plots for the UIRT oral composite and the MIRT estimates. As a point of comparison, the point plot for the UIRT oral composite

and the UIRT average was presented in figure 3.11. Similar plots for the Upper Elementary population were presented in Figure 3.12 to 3.14.

These figures suggest that for both populations, there seems to be a linear relationship between the UIRT oral composite and the first latent ability measured by the MIRT model while there seems to be a non-linear relationship between the UIRT oral composite estimates and second latent ability measured by the MIRT model.

Figure 3.9

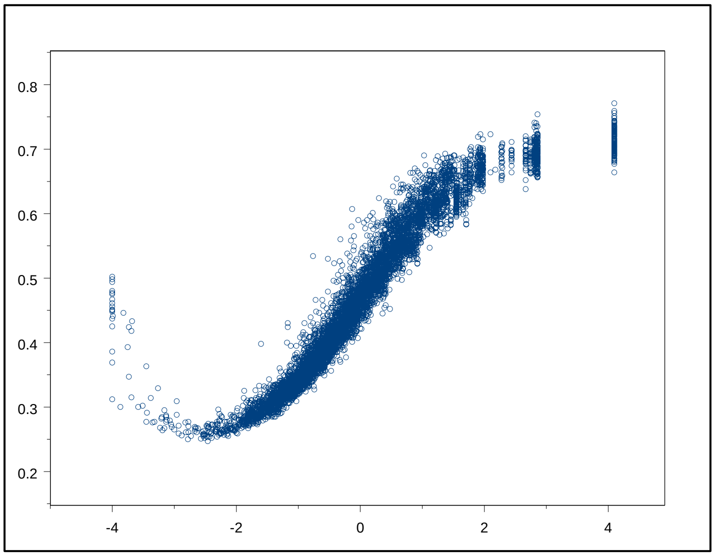
Point Plot of Oral Composite versus the MBIRT θ_1, Lower Elementary



Oral Composite

Fig 3.10

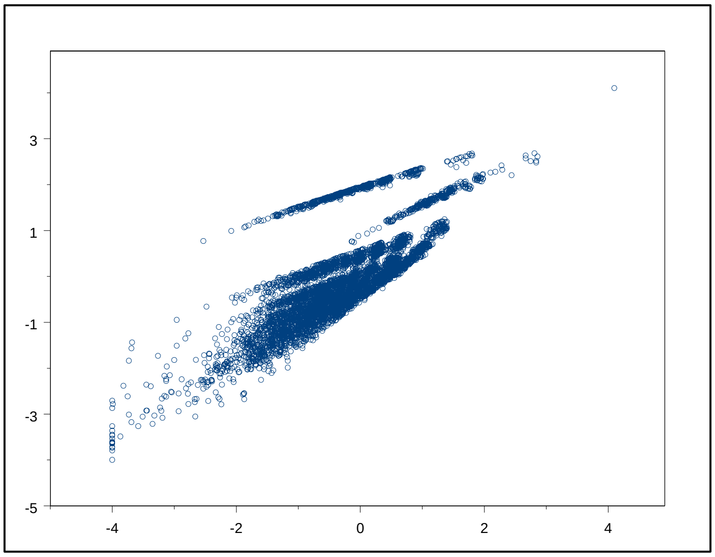
Point Plot of Oral Composite versus the MBIRT θ_2, Lower Elementary



Oral Composite

Fig 3.11

Point Plot of Oral Composite versus the Average UIRT Estimates, Lower Elementary



Oral Composite

Fig 3.12

Point Plot of Oral Composite versus the MBIRT θ_1, Upper Elementary

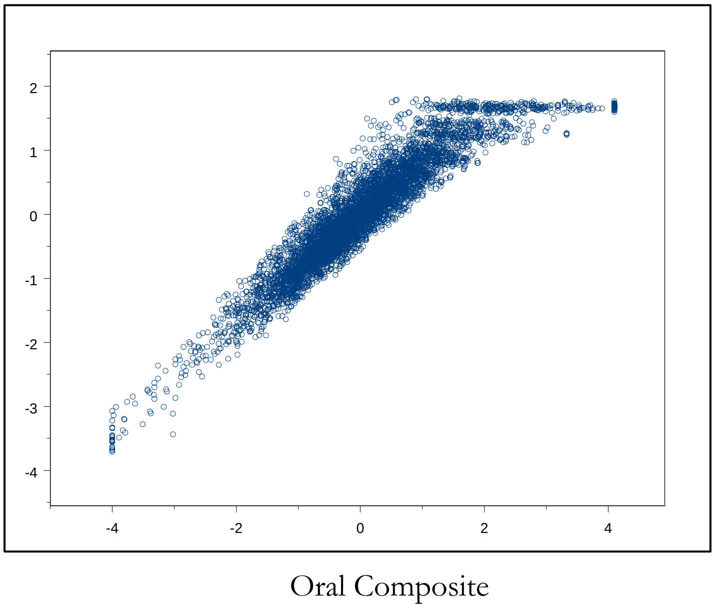
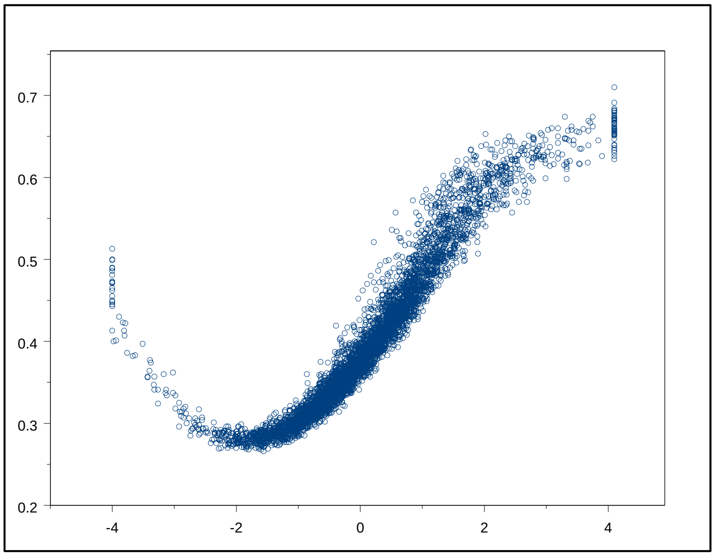


Fig 3. 13

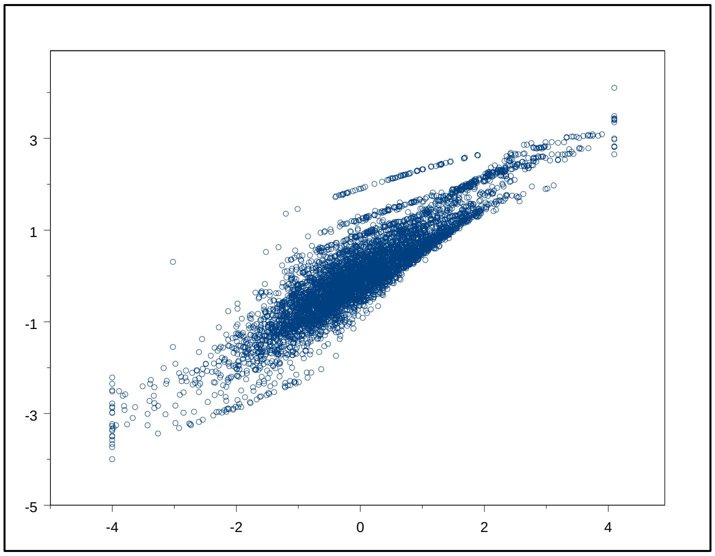
Point Plot of Oral Composite versus the MBIRT θ_2, Upper Elementary



Oral Composite

Fig 3.14

Point Plot of Oral Composite versus the Average UIRT Estimates, Upper Elementary



Oral Composite

DISCUSSIONS

Since the most common practice in designing English Language Assessment tests is to ensure that each subtest is one-dimensional. At the same time, it is also necessary to create composite score that is a function of two relatively uni-dimensional scores. The most common approach of creating composite score is by taking the simple average of two scores. The other approach is to conduct a concurrent calibration of two relatively uni-dimensional scales. The main issue with this approach is that it is not necessarily guaranteed that the composite domains which consist of more than one subtest are one-dimensional. Thus it is likely that multidimensionality exists in the composite domains. When multidimensionality exists, multidimensional item response theory has been shown to produce more accurate and efficient parameter estimates (Reckase, 1985). Thus, the use of multidimensional item response theory in composite score creation may provide better composite estimates.

This paper demonstrates that the multidimensional item response model can be used as an alternative way of creating composite scores using the LAS Link Speaking and Listening subtests. A factor analysis of for performance on the Oral scale which is a combination of Speaking and Listening subtests show that there are important, although subtle, second dimensions present in the test.

Results from the multidimensional analyses indicated that the MIRT procedure can be successfully used in modeling secondary ability dimensions on Oral ability assessment for the elementary schools populations. For the Upper Elementary population, both the Akaike information criterion and the chi-square difference test identified that a two-dimensional MIRT model was the best fitting model in representing the test structure of the Oral scale. For the Lower Elementary population, the Akaike information criterion identified that a one-dimensional IRT model was the best fitting model while the Chi-Square difference test identified that a two-dimensional MIRT model was the best fitting model. Such inconsistencies in the results provided by two different fit statistics warrants further studies using simulation studies.

In the psychometric literature, there seems to be little consensus on any best methods for determining and interpreting the dimensionality of the latent trait space with respective to both psychometric criterion (Stout 1987.) The problem is further compounded when an exploratory approach such as the one used in this paper was used to determine the structure of the latent space. Under exploratory approach, the dimensionality problem is compounded by the need to resolve issues such as factor invariance and rotational indeterminacy.

Recent research has suggested that for tests of less than 100 items, MIRT model may not be as capable of discriminating among examinees traits as uni-dimensional models (Davey & Hirsch, 1990.) This problem would be due to the increase in the parameterization which complicates the identifiably of the additional parameters needed to be estimated for MIRT model. These types of problems have hindered the development of practical applications involving MIRT models. In many instances, multidimensionality is ignored completely in favor of a less complex uni-dimensional model.

Despite of limitations mentioned above, this study shows that it is possible to make use of available uni-dimensional Speaking and Listening subtests to obtain multidimensional composite score that better presents the latent dimension underlies the Oral scale.

It has to be noted that the three different types of composite scores: the average of Speaking and Listening scores, the concurrent calibration of Oral items, and the proposed multidimensional scores are making different assumptions and each has its own limitations. In situations when there is a clear a priori theoretical belief about the latent structure of the ability, a particular model would be adopted based on this belief. However, in most practical testing situation when the test dimension is not clear, finding the most adequate model by comparing several candidate models provide empirical justifications of using one particular model.

REFERENCES

Ackerman, T. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18,* 257-275.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov. & B. F. Csaki (Eds.) *Second Internal Symposium on Information Theory*. Budapest, Academiai Kiado.

CTB/McGraw-Hill, 2006. *Technical Manual for LAS Links K-12 Assessment*. Monterey, CA: Author.

Davey, T., & Hirsch, T. M. (1990). *Examinee Discrimination as a measure of test data dimensionality*. Paper Presented at the annul meeting the Psychometric Society, Princeton, NJ.

Gulliksen, H. (1950). Theory of Mental Tests. New York: Wiley.

Haberman, J. S. (1977). Log-linear models and frequency tables with small expected cell counts. *Annuals of Statistics, 5,* 1148-1169.

Hale, G.A. Stansfield, C.W., Rock, D.A., Hicks, M.M., Butler, F.A., & Oller, J. W. (1988). *Multiple-choice cloze items and the Test of English as a Foreign Language* (TOEFL Research Report, No. 26). Princeton, N. J.: Educational Testing Service.

Harris B. Tetrachoric correlation coefficient. In Kotz L, Johnson NL (Eds.), *Encyclopedia of statistical sciences. Vol. 9 (pp. 223-225)*. New York: Wiley, 1988.

Kane, M., & Case, M. (2004). The reliability and validity of weighted composite scores. Applied Measurement in Education, 17, 221-240.

Kolenikov, S. POLYCHORIC: A Stata routine for estimation of polychoric correlation matrices. Computer program documentation, 2004. Kolenikov, K., and Angeles, G. (2004). The Use of Discrete Data in Principal Component Analysis with Applications to Socio-Economic Indices. CPC/MEASURE Working paper No. WP-04-85.

Loehlin JC. Latent variable models, 3rd ed. Lawrence Erlbaum, 1999.

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

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

Patz, R., & Yao, L. (2003). *Hierarchical and multidimsional models for vertical scale*. Paper Presented at the annual meeting of national council on Measurement in Education, Chicago.

Pearson K. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A, 1901, 200, 1-66.*

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

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

Rudner, L. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice*, *20(1)*, 16-19.

Stout, W. F. (1987) A nonparametric approach for assessing latent trait dimensionality. Psychometrika, 52, 79-98.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-118.

Yao, L. (2004). Baysian multivariate item response theory and BMIRT software. Paper presented at the Joint Statistical Meetings, Toronto, Canada.

Yao, L. & Boughton, K. A. (2005). Multidimensional parameter recovery: Markov chain Monte Carlo versus NOHARM. *Manuscript under review*.

Yao, L. & Schwartz, R. D. (2005). A Multidimensional Partial Credit Model with Associated Items and Test Statistics: An Application to Mixed-Format Tests. *Applied Psychological Measurement*, *30*, 1-25.

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