The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners

Educational and Psychological
Measurement

2015, Vol. 75(5) 805–825

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414554416

epm.sagepub.com

Mark D. Reckase¹ and Jing-Ru Xu¹

Abstract

How to compute and report subscores for a test that was originally designed for reporting scores on a unidimensional scale has been a topic of interest in recent years. In the research reported here, we describe an application of multidimensional item response theory to identify a subscore structure in a test designed for reporting results using a unidimensional scale. This research also dealt with the problem of planned missing data due to low levels of item overlap among multiple test forms. Furthermore, we provided evidence for the generalizability of the multidimensional structure using multiple forms of the same test. We also compared the subscores from multiple groups to show the usefulness of the subscores. The research provides evidence that subscores can be identified and produced to provide useful information about different constructs for multiple examinee groups even though the test data were well fit by a unidimensional model.

Keywords

subscores, multidimensional item response theory (MIRT), dimensionality, missing data

There is an increasing interest in subscores in educational testing because subscores have potential benefits in remedial and instructional applications (Sinharay, Puhan, & Haberman, 2011). Policy makers, college and university admissions officers, school district administrators, educators, and test takers all want subscores to help them make decisions for both admission and diagnosis purposes (Monaghan, 2006).

¹Michigan State University, East Lansing, MI, USA

Corresponding Author:

Mark D. Reckase, Michigan State University, 461 Erickson Hall, East Lansing, MI 48824, USA.

Email: reckase@msu.edu

The National Research Council (2001) report “Knowing What Students Know” emphasizes that the goal of assessment is to provide useful information for examinees’ knowledge, skills, and abilities. Also, the U.S. Government’s No Child Left Behind Act of 2001 requires that students should receive diagnostic reports that allow teachers to address specific diagnostic needs. Subscores can be used to identify such particular information for examinees and to report diagnostic analyses for teachers (Sinharay et al., 2011).

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

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

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

According to the Standard 5.12 of the *Standards of Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education,

1999), “Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established.” Also, the Standard 1.12 further clarifies:

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

The requirements implied by these standards can be summarized by the following question: Can the subscore structure defined by the test specifications be supported with empirical evidence about the number of domains, the relationship of item types to domains, and the number of items within each domain? If the dimensional structure of the test is not appropriately identified, then reporting subscores based on such fixed domains may not be meaningful.

Second, almost no research focuses on the dimensional structure of the test across multiple forms. In most research, the data sets are either from simulations or from a single set of real test data. Using multiple forms to support the inferences about the dimensional structure of the test for reporting subscores is very important for showing the generalizability of the results. This is especially important when subscores are reported for diagnostic purposes for multiple groups of examinees with differences in demographic and language background.

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

Furthermore, in most research the subscores are considered for individual-level reporting instead of an aggregated-level such as for multiple groups of examinees with different educational and demographic backgrounds. In practical educational services, representing the group differences using subscores has many meaningful applications, especially for teachers and school administration officers. For example, for international English language tests such as TOEFL and IELTS, providing subscores for different language groups according to examinees’ linguistic and cultural backgrounds will help not only examinees but also language teachers, language educators, language institutions, and linguistics researchers to improve and analyze students’ English learning in an informative way.

Given the issues identified in the research literature on subscores, three research questions were identified as the focus for the research reported here. First, can MIRT methods be used to identify a reasonable subscore structure for a large-scale test that is well fit by a unidimensional model? This question is addressed in the context of

real test data with multiple test forms and the associated problems of missing data. Second, is there evidence that the multidimensional subscore structure generalizes over multiple forms from the same test? Finally, if a subscore structure is identified across multiple test forms, do reports of the subscores provide meaningful information about multiple groups of examinees with different characteristics.

The data for this research came from a relatively new test of English for those who have different first languages—the *Pearson Test of English Academic* (Pearson Longman, 2010). This test was selected for analysis because it has thorough coverage of the components of English language and item response data were available from individuals from a number of different countries and language backgrounds. There are some complexities in the use of the data from this program, however. The *Pearson Test of English Academic* has many different test forms and a complex pattern of common items between forms. This makes the analyses of the data from this program challenging. However, through careful analysis, these challenges were overcome and the data were used to address the following specific research questions:

1. How many distinct dimensions are needed to accurately describe the relationships between the test items for the current heterogeneous sample of examinees? In particular, is more than one dimension needed?
2. If more than one dimension is needed to represent the relationships among the test items for the current sample of examinees, are there clusters of items that are sensitive to distinct combinations of skills and knowledge and are these clusters related to known constructs of language performance?
3. If meaningful clusters can be identified, are they specific to one form of the test or do similar clusters appear for more than one form? That is, do multiple forms replicate the complex structure of the language constructs?
4. If replicable clusters can be identified, do scores on the sets of items in the clusters produce meaningful subscores that show useful differences in English skills for multiple groups with different first languages?

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

Method

Data Description and Data Analysis Procedure

The purpose of the test used in this research is to measure nonnative English speakers' ability to function in an academic environment where instruction is provided in English. Specifically, this test assesses proficiency in listening, reading, speaking,

and writing English. It uses 20 different item types that assess different communicative skills, enabling skills, and other traits. Detailed descriptions of the item types are provided in Pearson Longman (2010).

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

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

Of the 164 test forms, 4 were found to have sufficient data for the multidimensional analyses. The minimum sample size for these four forms was 432. Thus, the analysis data consisted of five data sets. The first data set is the 100 items with highest frequencies of use. This was used to obtain results that could generalize across all test forms. The second to fifth data sets are from the four test forms with highest frequencies of administration. These were used to confirm the results from the 100 most frequently used items and to check the consistency of findings across forms.

Table 1 shows the distributions of item types for a test form. Table 2 provides the number of examinees and the number of items for the five data sets. Table 3 shows the number of common items between pairs of the five analysis data files.

Three Major Examinee Groups. For each form, the three major language groups that had the largest sample sizes were identified to determine if there were useful subscore differences across such groups. The full data set had examinees who reported 111 different home languages. English, Urdu, and Chinese-Mandarin were the three most frequent languages. They accounted for 46% of home languages. For each of the four forms with highest frequencies, the three language groups that had largest sample sizes were identified for the further analysis. They were the Asian group

Table 1. Item Types and Content Distribution for One Form of the Test.

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

Table 2. Number of Examinees and Number of Items for the Five Analysis Data Sets.

Data sets	Number of examinees	Number of items
Data Set 1 100 items with highest frequencies	36,938	100
Data Set 2 F1	448	65
Data Set 3 F2	438	53
Data Set 4 F3	437	69
Data Set 5 F4	432	66

Table 3. Common Items Between Pairs of the 100 Items and Four Test Forms.

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

Table 4. Sample Size for Each Language Group by Test Form.

Total of Three Language Groups		ENG	URD	ASN	ASN*			
					CMN	JPN	KOR	YUE
F1	221	121	42	58	29	7	11	11
F2	224	128	48	48	29	4	9	6
F3	249	134	69	46	23	4	7	12
F4	220	122	54	44	30	2	7	5

Note. ASN* shows the distribution of the ASN category composed of CMN, JPN, KOR, and YUE. ASN = Asian; ENG = English; URD = Urdu; CMN = Chinese-Mandarin; JPN = Japanese, KOR = Korean; YUE = Chinese-Cantonese.

(ASN), the English group (ENG), and the Urdu group (URD). The Asian group includes Chinese-Mandarin (CMN), Japanese (JPN), Korean (KOR) and Chinese-Cantonese (YUE).

The English group has the largest sample across all the four forms—almost twice the size of either the Asian or Urdu group. Table 4 presents the sample size for each language group taking each test form.

Dimensionality Analysis

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

Analysis of Most Frequently Used Items

Parallel analysis. Parallel analysis is one approach for determining the number of dimensions needed to describe the interactions between variables in a data set (Reckase, 2009). In this study, the first step consisted of computing the eigenvalues from the interitem correlations among the 100 most frequently used items. Correlations were computed using pairwise deletion to account for missing data across forms resulting in a different sample size for the correlations of each pair of items. This can influence the results of the dimensional analysis. Therefore, when the comparison random data sets were generated with the same proportion of item scores for each item, individual item scores were removed to exactly match the pattern of missing values in the real data set. Then, the eigenvalues from the generated data set were extracted from the interitem correlations. This process was replicated 100 times to yield distributions of the eigenvalues from the randomly generated data sets.

Because of the pattern of missing data, there were cases where a correlation could not be computed between a pair of items. Without the full correlation matrix, the eigenvalues could not be computed so the missing values were imputed by predicting

Figure 1. (A) Plot of the eigenvalues for the real data and 100 replications of random data. (B) Magnified plot of the number of eigenvalues larger than random data.

the missing values using the data from all the other columns of the correlation matrix using multiple linear regression. Because exactly the same process was used for the real data and the randomly generated data, any artifacts caused by the imputation and the pattern of missingness would be present in both types of data sets. Figure 1A shows a plot of the magnitude of the eigenvalues from the real data and those from the 100 replications of the randomly generated data. Because the eigenvalues from the generated data showed little variation, the results where the curves cross are

magnified and presented in Figure 1B so the number of eigenvalues from the real data that are greater than those from the random data can be identified.

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

Exploratory factor analysis. In order to check the viability of using seven dimensions to analyze the data, exploratory factor analyses were run on the 100-item data set using Mplus (Muthén & Muthén, 2005) specifying successively one to seven dimensions. The results showed that the seven-dimensional solution gave the best combination of distinctly defined factors with multiple item loadings and no factors that appeared to represent error from overfactoring. The rotated factor loadings, the estimated residual variances, the factor correlations all supported the seven-factor structure. However, the missing data made it impossible to compute meaningful fit statistics. To further check the meaningfulness of the seven dimensional structure, cluster analysis procedures were used to determine if sets of items had theoretically supportable connections to the content structure of the tests.

Hierarchical cluster analysis. Within the context of multidimensional item response theory, hierarchical cluster analysis is an approach for identifying sets of items that are best at measuring the same combination of skills and knowledge. There are two steps in the cluster analysis procedure. The first step is to select a method to measure the similarity between items. The second step is to sort the items that share similarities into clusters (Reckase, 2009). Kim (2001) showed that hierarchical cluster analysis (especially the average method) can recover item clusters that were defined in a simulation study.

In this study, the factor loadings from the seven-factor solution were used as the item discrimination parameters for the multidimensional item response theory model. These parameters were used to compute the angle between the directions of best measurement for pairs of items. The average distance algorithm in the hierarchical clustering routine within MatLab was used for the clustering using the angles as the similarity measure. Figure 2 shows the dendrogram for the 100-item set based on the seven-dimensional solution from the exploratory factor analysis.

The cluster results shown in Figure 2 indicate that six distinct clusters were identified through the analysis of the 100-item data set. Moreover, among these six clusters, five distinct clusters consist of unique collections of item types and one cluster is composed of a mix of five different item types. The five major clusters were labeled according to the conceptual representation of factors in the language ability domain defined by Carroll (1993). They are (1) Word Recognition and Pronunciation, (2) Reading Comprehension, (3) Spelling and Phonetic Coding, (4) Oral Production, and

Figure 2. Dendrogram for cluster analysis of 100 items with highest frequency.

(5) Listening and Oral Production. The sixth cluster was labeled as writing and listening comprehension.

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

The reference composites were computed for each of the clusters of items identified by the cluster analysis procedure. The reference composites represent the distinct subscores that can be supported by the set of items—in this case the 100 items with the highest frequency of use. One way of computing the subscores is to project the estimates of locations of the examinees in the seven-dimensional space onto these reference composite lines. See Reckase (2009) for the details of the projection method.

To compute the reference composite for all the items within each cluster, the $\mathbf{a}'\mathbf{a}$ -matrix was obtained and decomposed into eigenvalues and eigenvectors. Table 5 gives the angles in degrees between each reference composite line and the coordinate axes in seven-dimensional space for each cluster of the 100-item set. Table 6

Table 5. Angles Between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Six Clusters.

Dimension-axes	Oral Production	Listening and Oral Production	Word Recognition and Pronunciation	Spelling and Phonetic Coding	Reading Comprehension	Listening and Writing and Comprehension
1	5.07	81.96	88.69	89.34	88.72	88.54
2	88.11	88.13	2.41	87.88	89.27	87.89
3	89.50	89.10	89.70	87.05	6.46	87.43
4	89.22	87.67	89.46	5.58	85.81	84.57
5	85.50	8.92	89.04	87.86	89.57	86.04
6	89.05	87.76	89.42	86.41	85.36	7.66
7	89.75	89.94	88.44	89.89	89.71	89.75

Table 6. Angles Between Each Pair of Reference Composites for Six Clusters.

	Oral Production	Listening and Oral Production	Word Recognition and Pronunciation	Spelling and Phonetic Coding	Reading Comprehension	Listening and Writing and Comprehension
Oral Production	0	77.33	86.70	88.24	88.04	87.13
Listening and Oral Production	77.33	0	86.95	85.22	88.13	83.33
Word Recognition and Pronunciation	86.70	86.95	0	87.23	88.85	87.14
Spelling and Phonetic Coding	88.24	85.22	87.23	0	82.53	80.63
Reading Comprehension	88.04	88.13	88.85	82.53	0	82.34
Writing and Listening Comprehension	87.13	83.33	87.14	80.63	82.34	0

presents the angles between each pair of reference composites in the seven-dimensional space.

The results in the table show that the reference composite lines tend to match one of the coordinate axes in the multidimensional θ -space. For example, the Oral Production cluster has a reference composite line that is very close to the Dimension 1 coordinate axis—its angle with the axis is only 5° . The same relationship can be observed for the reference composites of the other clusters as well. Thus, each cluster defines a unique dimension corresponding to a particular language ability and aligns with a coordinate axis in the solution. Furthermore, the angles between each pair of reference composites show that these six clusters are almost orthogonal to each other in this solution indicating that these six clusters represent six distinct abilities. Note, however, that this result does not imply that the constructs measured by the sets of items are uncorrelated. It only means that the solution has been rotated so that the coordinate axes match each of these reference composites. When scores on sets of items from the clusters are correlated and corrected for attenuation, some of the correlations are in the .90s. These high correlations are what allow a unidimensional item response theory model to fit the full set of items for a test form. The correlations for the clusters are presented later in this article after the discussion of the analysis of individual test forms.

Despite these high correlations between subscores, these results show that there are distinct statistically identified scores based on the set of items in the 100 most frequently used items. However, these constructs may not have substantive meaning, although they are connected to the item types selected to measure specific language constructs. Based on these results, it is clear that there exist multiple dimensions in the data that may be related to important language constructs.

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

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

Consistency of dimension structure. To determine the consistency of structure across forms and the most frequently used 100 items, the common language constructs across four test forms were identified. Figure 3 shows a comparison between clusters identified within each form and the clusters extracted from the 100 items with highest frequencies. The number of clusters identified for the 100 items with highest frequencies (100), F1, F2, F3, and F4 are 7, 8, 6, 7, and 6, respectively. The corresponding

Table 7. Correlations Between Eight Clusters From F1.

	OP	WRP	CL	RW	PhP	LR	LW	CLR
OP	1	0.67	0.52	0.26	0.57	0.31	0.42	0.34
WRP	0.72	1	0.52	0.37	0.64	0.34	0.42	0.32
CL	0.61	0.60	1	0.46	0.75	0.58	0.64	0.46
RW	0.36	0.50	0.67	1	0.49	0.45	0.54	0.41
PhP	0.64	0.71	0.91	0.70	1	0.55	0.60	0.47
LR	0.39	0.41	0.76	0.70	0.70	1	0.52	0.35
LW	0.53	0.52	0.87	0.86	0.79	0.75	1	0.33
CLR	0.65	0.59	0.95	0.98	0.92	0.74	0.72	1

Table 8. Eight Cluster Names in F1.

OP	WRP	CL	RW
Oral Production	Word Recognition, Pronunciation	Communication, Listening	Reading, Writing
PhP	LR	LW	CLR
Phonetic Coding, Pronunciation	Listening, Reading	Listening, Writing	Communication, Listening, Reading

cluster names are also alphabetically sorted in Figure 3. In the figure, the black squares indicate that clusters from the different item sets share exactly the same language constructs whereas the grey boxes indicate that only part of the constructs are the same between two clusters. Thus, F1 and F3 have very similar dimension structures. Most of the forms share some of the constructs with the 100-item set, but not all of them. That is not surprising because the 100 most frequently used items did not include all the item types. It appears forms F2, F3, and F4 show strong multidimensional parallelism and share some of the constructs with the 100-item set.

Reliabilities of the subscores and correlations between the clusters. For each of the clusters identified for one of the test forms, subscores were obtained on the sets of items. Raw correlations, correlations corrected for attenuation, and reliability for observed subscores based on the number of clusters identified for each test form were computed. Since the dimension structure across different forms proves to be consistent, the analysis results of the correlations and reliabilities for F1, F2, F3, and F4 are very similar. F1 was selected for illustration purpose.

Tables 7 and 8 give the correlations between the scores on the item clusters for F1. The raw correlation is above the diagonal including the diagonal and the correlations corrected for attenuation are given below the diagonal. Table 9 presents the Cronbach alpha reliability of the subscores based on eight clusters for F1. Most of the clusters

Table 9. Reliability for Eight Clusters in F1.

Number of cluster	Cluster names	Number of items	Reliability
1	OP	9	.92
2	WRP	5	.96
3	CL	16	.79
4	RW	6	.57
5	PhP	14	.85
6	LR	7	.72
7	LW	4	.68
8	CLR	4	.30

give scores that are distinct from the others, but some pairs, such as the RW and CLR clusters (both measure reading skill), CLR and CL clusters (both measure communication and listening skill), have very high correlations corrected for attenuation. The reliabilities for most clusters are within a reasonable range, but for the clusters that have a small number of items (less than five items) the reliabilities are much lower (about .3 for Cluster 8). These are typical results obtained for other test forms.

Evidence of Usefulness of Identified Item Clusters

For each of the clusters identified for one of the test forms, subscores were obtained on the sets of items and then standardized to have a mean of 0.0 and a standard deviation of 1.0 for the full sample of examinees. Because one major research purpose is to compare the subscores among different examinee groups instead of among each individual examinee, mean subscores were computed for each of the three major language groups who took the test forms. If the data were consistent with a unidimensional model, it would be expected that the profiles of performance would differ in level based on the overall mean performance, but the profiles would be parallel to each other. However, for each form, the analyses showed that the performance profiles for three language groups were not parallel to each other. The differences of the subscore profiles for three language groups on F3 are shown in Figure 4 for illustration purposes. Table 10 lists the corresponding means and standard deviations of subscores for three language groups on F3.

A multivariate analysis of variance and Tukey’s post hoc tests were performed to test for significant differences of construct means across the three language groups for the four forms. The results showed that there is an overall significant difference of the construct means across different language groups. However, the Tukey’s post hoc tests indicate that mean scores between some group pairs are not significant. In Figure 4, the pairs that are not significantly different are enclosed in ovals.

The results showed that English group has much higher scores on most clusters than the Asian group and the Urdu group, but it has larger variation in scores on the clusters than the Asian and Urdu groups. The Urdu group has the smallest variation

Table 10. Means and Standard Deviations of Subscores on Clusters of Items for the Three Language Groups on F3.

Clusters	ASN	ENG	URD
1	-0.59 (1.10)	0.10 (1.00)	0.29 (0.91)
2	-0.62 (0.77)	0.14 (1.02)	0.06 (1.07)
3	-0.55 (0.94)	0.27 (1.06)	0.03 (0.80)
4	-0.19 (0.84)	0.22 (1.18)	-0.10 (0.70)
5	-0.31 (0.97)	0.24 (1.05)	-0.07 (0.85)
6	-0.22 (0.94)	0.21 (1.09)	-0.08 (0.82)
7	-0.86 (0.99)	0.39 (0.96)	0.02 (0.95)

Note. ASN = Asian; ENG = English; URD = Urdu.

Figure 4. Profiles of mean performance on clusters of items for the three language groups on F3.

in scores whereas the Asian group has the lowest score for most clusters among three language groups. It is not surprising that the English group tends to perform best—it should. It is interesting, however, that the Urdu group has the highest scores on Oral Production with the smallest variation among the three language groups. This is typical of the results on the other forms.

Moreover, the profiles and Table 10 show that the Asian group is very similar to the Urdu group on Reading Comprehension, Oral Communication and Listening

Comprehension, but performs at a much lower level on the other constructs—especially Phonetic Coding and Pronunciation. These results show the potential value of reporting subscores by language group, or possibly by teaching approach, as a way of identifying areas that are in need of improvement.

Summary and Conclusions

Summary of the Results

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

The analysis of language group means suggests that reporting subscores based on different language domains provide meaningful information. The significant differences of the subscore profile patterns between different language groups are consistent with the characteristics of the different languages and the familiarity of those groups with English. Further analysis should be done on other language groups and on groups that have followed different instructional approaches to further show the value of subscore reporting.

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

In conclusion, this study explores the support for the validity of the multidimensional structure across multiple test forms when the test was originally designed for a unidimensional scoring procedure using the Rasch model. Through the analysis, we can support the use of subscores for reporting, especially to show the differences in performance for different first language groups. The analyses suggest that six to eight dimensions are needed to represent the constructs assessed by the different test forms. The analysis of Data Set 1—100 items with highest frequencies across all test forms—showed that the very distinct seven-dimensional solution was needed to

accurately describe the relationships between the test items and the current sample of examinees. The analyses of data sets from four test forms were consistent with the 100-item analyses, supporting six to eight dimensions, even though the samples were small. The correlations and reliabilities show that the clusters identified for each data set are sensitive to distinct combinations of skills and knowledge. Figure 3 shows there was a consistency of the dimension structure across five data sets, indicating that the language constructs can be replicated across multiple forms. Moreover, the performance profiles of the clusters for three language groups are not parallel to each other in all test forms. Therefore, the subscores on the sets of items in these clusters provide meaningful differences in English skills for examinee groups with different first languages.

Important Issues

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

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

The second issue is that large assessment programs use multiple test forms and the dimensional structure needs to be replicated over the forms. This means it is necessary to check whether the number of dimensions and the structure generalize over multiple test forms. The scores received from multiple test forms and administrations should be comparable (Wendler & Walker, 2006). Moreover, “Validity is a unitary concept . . . [it] refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (Messick, 1993, p. 13). Therefore, for large-scale testing programs, in order to provide the validity evidence supporting the number and types of subscores for different groups of examinees, we need to provide evidence for a consistent dimensional structure across multiple test forms.

The third issue is whether it is useful to report subscores when they are fairly highly correlated in a population made up of multiple groups of examinees. The results reported here indicate that there are meaningful differences in profiles for subgroups despite the high intercorrelations of constructs. It may be that correlations are

high enough that score reporting for individual examinees is not supported. However, the subscores can still provide meaningful diagnostic and instructional information for examinees and users of subscores such as instructors, education administrators, and educational researchers, based on group reports.

In this study, subscores are reported for groups of examinees. The pattern of subscore means for different language groups presents both differences and commonality of characteristics related to linguistic factors regarding different home languages. Such differences and similarities will provide useful and meaningful comparison between language groups having different linguistic characteristics.

Implications for Future Study

Currently, the multiple forms of the test used in this research are equated using unidimensional IRT linking. This works well because the multiple constructs identified through the analyses reported here were fairly highly intercorrelated. Yet the analyses support the reporting of multiple subscores. To show the consistency of the dimension structure across multiple forms, we compared different language cognitive factors that were represented by each item type using the conceptual representation of factors in the language ability domain defined by Carroll (1993). For future study, it would be more efficient and accurate to confirm the dimension structure across multiple forms when these multiple forms are equated using MIRT linking procedures.

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

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Mark D. Reckase participates on a technical advisory committee for Pearson-Edexcel.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a research grant from Pearson-Edexcel. Pearson-Edexcel also made the test data available for analysis for this project.

References

- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51.
- Ackerman, T., & Shu, Z. (2009, April). *Using confirmatory MIRT modeling to provide diagnostic information in large scale assessment. Paper presented at the meeting of the National Council of Measurement in Education*, San Diego, CA.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University Press.
- Haberman, S. J. (2008a). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204-229.
- Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, 75, 209-227.
- Harris, D. J., & Hanson, B. A. (1991, March). *Methods of examining the usefulness of subscores. Paper presented at the meeting of the National Council on Measurement in Education*, Chicago, IL.
- Kim, J. L. (2001). *Proximity measures and cluster analyses in multidimensional item response theory* (Unpublished doctoral dissertation). Michigan State University, East Lansing.
- Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12, 1-11.
- Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Phoenix, AZ: Oryx Press.
- Monaghan, W. (2006). *The facts about subscores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections4.pdf
- Muthén, L. K., & Muthén, B. O. (2005). *Mplus user's guide* (Version 3). Los Angeles, CA: Muthén & Muthén.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- Pearson Longman. (2010). *The official guide to PTE: Pearson Test of English Academic*. Hong Kong SAR: Pearson Longman Asia ELT.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47, 150-174.
- Sinharay, S., Haberman, S. J., & Puhán, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26(4), 21-28.
- Sinharay, S., Puhán, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, 30(3), 29-40.
- Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, 23, 63-86.

- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287-307.
- Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., & . . . Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.
- Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: the effect of latent space misspecification on the application of IRT* (Research Report MW: 6-24-85). Iowa City, IA, University of Iowa.
- Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.
- Wendler, C. L., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445-467). Mahwah, NJ: Lawrence Erlbaum.
- Yao, L. H., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31, 83-105.
- Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64, 213-249.