Research Note: Using Multidimensional Item Response Theory to Link Multiple Test Forms and Report Subscores for Pearson Test of English Academic

Jing-Ru Xu
Pearson VUE, Chicago, U.S.A.
jing-ru.xu@pearson.com

Mark Reckase
Michigan State University, U.S.A.
reckase@msu.edu

June 2016

1. Introduction

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

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. Thus, the distinct domains defining subscores may not be clearly defined in a technical psychometric sense. Also, little information may be provided to show there are enough items in each domain to support reporting useful scores. Moreover, it may not be clear why particular types of items are grouped together within each domain. Finally, there are few discussions of how to link and equate test forms when reporting subscores.

In order to fill in the above gaps and to explore solutions to the questions, this research applied multidimensional item response theory to report subscores for a large-scale international English language test – the Pearson Test of English Academic (PTEA). Different statistical and psychometric methods were used to analyze the dimension structure, the clusters for reporting subscores, and to link individual test forms to provide comparable and reliable subscores. This research is a follow-up study of Reckase and Xu. (2014), which demonstrates a subscore structure for PTEA.

2. Data Description

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

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

Of the 164 test forms, four were found to have sufficient data for the multidimensional analyses. The minimum sample size for the 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 provides the number of examinees and items within each data set.

Table 1: Number of examinees and number of items for the five analysis data sets

Data Sets	Number of Examinees	Number of Items
Dataset 1 F100	36938	100
Dataset 2 Form F1	448	65
Dataset 3 Form F2	438	53
Dataset 4 Form F3	437	69
Dataset 5 Form F4	432	66

3. Methodologies and Results

Simulated data sets were designed according to the real data sets formats to replicate the exact missingness pattern and the item scores. Furthermore, both dichotomous and polytomous items with low frequencies for particular score categories were recoded so that item parameters could be calibrated using multidimensional item response models. Also, results from different software analysis packages were compared to validate the efficiency and accuracy for further analyses using both real and simulated data sets with missingness recovery skills.

3.1 Analyses of the data structure

Parallel analyses were applied to the eigenvalues among the 100 most frequently used items to identify the dimensional structure across all data sets. The whole procedure was achieved pair-wise to account for the missing data across different forms. For comparison purpose, random data sets were generated with the same proportion of item scores for each item, individual item scores were removed to exactly match with the pattern of missing values in the real data sets. Then, the eigenvalues from the generated data set were extracted from the inter-item correlations. This process was replicated 100 times to yield distributions of the eigenvalues from the randomly generated data sets. The results showed that either a seven or eight dimensional structure can be used to explain the relations among the data sets. Figure 1 presents the eigenvalues for the real PTEA data and 100 replications of random simulated data.

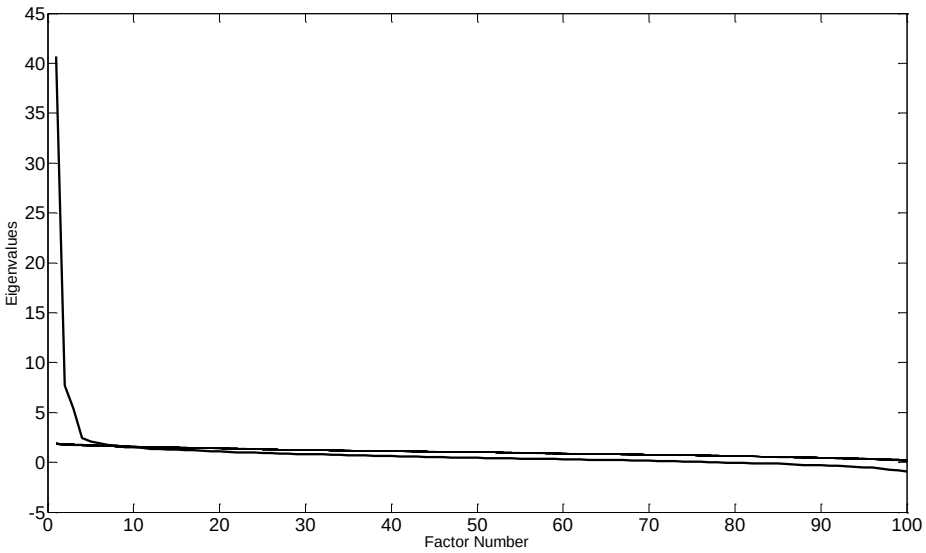

Figure 1: Plot of the eigenvalues for the real data and 100 replications of random data

The combination of exploratory and confirmatory factor analyses further confirmed the seven-dimensional structure of the 100 most frequently used items. The discrimination item parameters for the multidimensional item response theory models were calibrated and applied to run hierarchical clustering

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

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 (Wang, 1986). This scale is the one that would be obtained if the items were analyzed using a unidimensional item response theory model. The reference composites were computed for each of the clusters of items identified by the cluster analysis procedure. They represent the distinct subscores that can be supported by the set of items. Table 2 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 2: Angles between the Reference Composites and the Coordinate Axes in Seven-Dimensional Space for Six Clusters in Form F100

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

PhonCo: Phonetic Coding

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

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. The number of clusters identified for the 100 items with highest frequencies (Form F100), Forms F1, F2, F3, and F4 are 6, 6, 6, 6, 6, and 8, respectively. Figure 2 shows the eight clusters

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

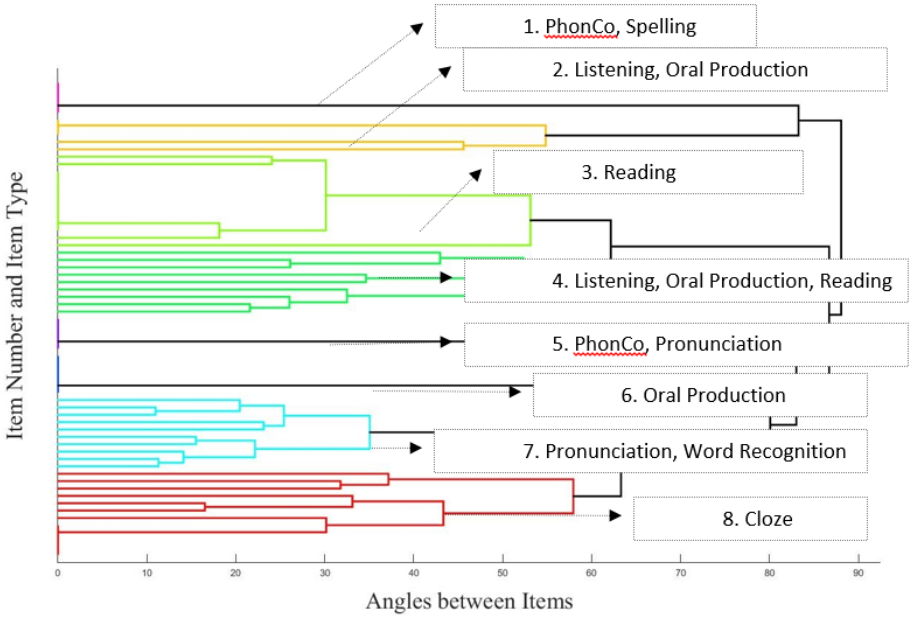

Figure 2. Clusters from Form F4

3.2 Linking of multiple test forms

We selected Form F100 as the base form, since it was extracted based on the sorted data with highest frequencies across all forms and all examinees. Forms F100 and Form F1 have 16 common items, which were distributed on six common subdomains C1 – C6. Form F1 has 65 items in total. Among those uncommon items, which are 47 items = 65 – 16, we found the items that have item types exactly the same as those in the six common constructs. We ran both exploratory and confirmatory factor analyses to determine the loadings of the uncommon items on subdomains. However, the estimation of item parameters for these common items using multidimensional item response model did not converge well. It might be due to the fact that there were few common items between the two forms. It could also be due to the facts that the data set has a mix of dichotomous and polytomous items. Table 3 lists the number of common items among five data sets.

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 3 shows that there is a small number of common items between individual forms and the base form – F100. Alternative methods were applied to link test forms with multiple subdomains.

We then applied the nonorthogonal Procrustes rotation method to obtain four different rotation matrices from each individual test form, respectively (Reckase, 2009). With the four rotation matrices, we were able to rotate subscores from each new form onto the base form. After obtaining these four rotation matrices, we post-multiplied these four matrices by the estimated abilities for examinees in each individual form to get the estimates onto common scales.

In all, there are 11 clusters in the common coordinate system including the clusters identified across all five data sets. The base form was built upon an 11-dimensional space. For each individual form, there was a corresponding 11-by-11 reference composite matrix with rows indicating the 11 clusters and columns the number or dimensions. These four matrices are the augmented matrices with additional dummy columns indicating the number of clusters added and the dummy rows representing the clusters not generated from a particular individual test form. Each cluster is represented by the corresponding reference composite using the cosine of the degree between the reference composite and the coordinate axis.

Table 4 shows the subscores after the rotation from Form F4 to the base form. Since there are 432 examinees, only rotated subscores of the first 10 examinees were provided. Clusters 2, 5 and 8 are not originally from Form F4. For clusters that do not belong to a particular form, the rotated subscores are not meaningful because they do not represent any constructs that the form was designed to measure. Therefore, we would not recommend to report the subscores for the clusters that were not identified in that particular individual form. In general, the values of subscores for these types of clusters can be computed when rotating subscores from new form back onto the base form, yet these subscores will not provide any meaningful interpretation of examinees' true abilities. They just represent the numbers in the mathematical calculation. In other words, they are mathematically meaningful, but psychometrically meaningless.

Table 4: The rotated subscores after non-Procrustes rotation for Form F4

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

4. Conclusions

The overall analyses showed there was a consistency of the dimension structure across five data sets, indicating the language constructs can be replicated across multiple forms. Therefore, the subscores on the sets of items in these clusters provide meaningful differences in English skills for PTEA.

The analysis of data set – 100 items with highest frequencies across all test forms showed a 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 the other four test forms were consistent with the 100-item analyses, supporting six to eight dimensions, even though the samples were small.

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

This research also applied an innovative linking method to transform each individual test form back onto the base form using nonorthogonal-Procrustes rotation according to the clusters identified in the dimensionality analyses.

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. Through the analyses, we can support the use of subscores for reporting. The analyses suggest that six to eight dimensions are needed to represent the constructs assessed by the different test forms. Subscores from different test forms can be linked to compare the differences among examinees' ability levels. It is a pioneer study that demonstrates how to report subscores across different test forms in a multidimensional structure.

References

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

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

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

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

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

Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.