ED 296 592 FL 017 475

AUTHOR Oltman, Philip K.; Stricker, Lawrence J.

TITLE How Native Language and Level of English Proficiency Affect the Structure of the Test of English as a Foreign Language (TOEFL).

INSTITUTION Educational Testing Service, Princeton, N.J.

PUB DATE Mar 88

NOTE 15p.; Paper presented at the Annual Colloquium on Language Testing Research (10th, Urbana, IL, March 5-7, 1988).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Analysis; *Construct Validity; *English (Second Language); Language Proficiency; *Languages; *Language Tests; *Standardized Tests; *Test Construction; Test Interpretation; Test Items

IDENTIFIERS *Test Of English As A Foreign Language

ABSTRACT

A study examined the relationship of native language and level of English proficiency to the structure of the Test of English as a Foreign Language (TOEFL). Using all of the information provided by various responses to the test's items (the four alternatives, omitted, and not reached), the items' interrelations were analyzed by three-way multidimensional scaling for samples of examinees systematically varying in native language and level of English proficiency. Four dimensions were identified: three corresponded to the sections of the test, and the fourth was an end-of-test phenomenon. The dimensions were predominantly defined by easy items and were most salient for low-scoring examinees. Native language had little influence on results. The major conclusions were that the TOEFL's construct validity is supported, the test's interpretation varies with the examinees' English proficiency, easy and difficult items differ in their potential for diagnosis and global screening, and the dimensionality of the TOEFL and of competence in English depends on the examinees' English proficiency. (Author/MSE)

How Native Language and Level of English Proficiency Affect the Structure of the Test of English as a Foreign Language (TOEFL)

Philip K. Oltman and Lawrence J. Stricker
Educational Testing Service

Paper presented at the
Tenth Annual Colloquium on Language Testing Research
Urbana, IL, March, 1988.

Address for correspondence:

Philip K. Oltman
Educational Testing Service
17-R
Princeton, NJ 08541

Abstract

The aim of this study was to appraise the effect of native language and level of English proficiency on the structure of the TOEFL$^{(R)}$. The interrelations among TOEFL items, using all of the information provided by the various responses to the items (the four alternatives, omitted, and not reached), were analyzed by three-way multidimensional scaling for samples of examinees systematically varying in native language and level of English proficiency. Four dimensions were identified: three corresponded to the sections of the test, and the fourth was an end-of-test phenomenon. The dimensions were predominantly defined by easy items and were most salient for low-scoring examinees. Native language had little influence on the results. Major conclusions were that the TOEFL's construct validity is supported, the test's interpretation varies with the examinees' English proficiency, easy and difficult items differ in their potential for diagnosis and global screening, and the dimensionality of the TOEFL and of competence in English depends on the examinees' English proficiency.

The Test of English as a Foreign Language (TOEFL; Educational Testing Service, 1985) consists of three sections, Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension, and provides scores for each section as well as a total score. The test is intended to assess the ability of nonnative speakers to understand spoken English, to comprehend reading materials, and to recognize correct structural, grammatical, and lexical usage.

Responses on the TOEFL may reflect both the influence of the examinees' native language and their level of English proficiency. Work thus far has not appraised the independent influences of native language and level of English proficiency on TOEFL performance, and these variables are confounded in most of this research.

The purpose of this study was to appraise the influence of examinees' native language and level of English proficiency on the structure of the TOEFL. More specifically, the aim was to assess the interrelations among TOEFL items for groups of examinees that systematically varied in native language and level of English proficiency, going beyond the usual right versus wrong scoring to use all the information provided by the various responses to the items.

Method

Examinees and Test Form

The data were drawn from the 53,169 examinees who took the TOEFL in the May 1985 international administration and had complete information. The form had 146 operational items. Twenty-one subsamples of examinees, comprising seven language groups (Arabic, Chinese, Greek, Japanese, Korean, Malay, and Spanish) and three levels of performance on the TOEFL (High--total scores on the TOEFL of 543 and above; Medium--scores of 483 to 540; and Low--scores of 480 and below) were randomly drawn from the total sample. All language groups with approximately 400 or more examinees at each of the three performance levels were included. (The three levels were determined by trichotomizing the score distribution for the total sample.) Each subsample consisted of 400 examinees, except for 397 in the low-scoring Greek subsample.

Analysis

For each of the 21 subsamples of examinees, a 146 x 146 matrix of symmetrical tau coefficients (Goodman & Kruskal, 1954; Jacobson, 1976) among the items was computed. This coefficient, a measure of association between two nominal variables, indicates (on a scale from 0 to 1) the extent to which one variable is predictable from the other, and vice versa. In this analysis, each item is a nominal variable with six categories (the four alternatives, omitted, and not reached), and the tau between a pair of items is computed from the resulting 6 x 6 contingency table.

A three-way, metric multidimensional scaling analysis of the 21 tau matrices was carried out, using SINDSCAL (Pruzansky, 1975). Three-way scaling allows for variation among individuals in the salience of the dimensions (the "individuals" in the present application are the 21 subsamples).

The results of the scaling were subjected to two hierarchical cluster analyses (Ward, 1963), one on the 146 items and one on the 21 subsamples, to identify regions in the multidimensional space where items and subsamples formed groupings.

Results and Discussion

Item Dimensions and Clusters

Dimensions. Based on an examination of the variance accounted for and the interpretability of the dimensions, the four-dimension SINDSCAL solution was chosen.

Figure 1 presents the items plotted for each pair of dimensions. Dimension I was defined by the items (all relatively difficult) for the last two Reading Comprehension passages, located at the very end of the test. This dimension appears to reflect the degree to which items were omitted or not reached: for the total sample, the items' coordinates on Dimension I correlated -.94 with the proportion of omitted responses and -.88 with the proportion of not reached responses. Dimension II was defined by relatively easy Listening Comprehension items; Dimension III by easy Vocabulary and Reading Comprehension items; and Dimension IV by items in one of the two Reading Comprehension passages at the end of the test, at one pole, and easy Structure and Written Expression items, at the other pole.

Thus, the easier items in each section of the test defined three of the dimensions. An additional dimension was defined by difficult items associated with reading passages and appears to be an end-of-test phenomenon. The remaining items contributed little to the emergence of any of the dimensions.

See Figure 1

Clusters. The tree diagram for the cluster analysis of items appears in Figure 2. Seven clusters were interpretable. The clusters consisted of (a) Reading Comprehension items for the next to last passage in the test, (b) Reading Comprehension items from the last passage in the test, (c) easy Listening Comprehension items, (d) easy Vocabulary and Reading Comprehension items, (e) easy Structure and Written Expression items, (f) medium difficulty Structure and Written Expression items, and (g) difficult items scattered throughout the test--a kind of "general" cluster.

An examination of the locations of the clusters on the plots of dimensions in Figure 1 (the clusters are shown as ellipses, with the general cluster as a shaded ellipse), shows that the general cluster, unlike the others, was always located at the center of each of the plots, indicating that its items did not define any of the dimensions.

See Figure 2

Subsamples Clustered by Subject Weights

Subject weights. Figure 3 presents the subject weights for the language/level subsamples plotted for each pair of dimensions. The subject weights on all the dimensions were greater for the low-scoring subsamples, with the largest weights occurring for the low-scoring Arabic, Greek, Japanese, and Spanish subsamples on Dimension I. These results indicate that the dimensions were more salient for the low-scoring subsamples, and the end-of-test dimension (Dimension I) had greater salience for some of these subsamples, the only instance in which language group had an effect.

Clusters. Three subsample clusters were interpretable; they are shown (as ellipses) in Figure 3. They consisted of (a) low-scoring Arabic, Greek, Japanese, and Spanish; (b) low-scoring Chinese, Korean, and Malay, plus medium-scoring Spanish; and (c) the remaining subsamples (all medium- and high-scoring).

Inspection of the locations of these clusters on the plots of subject weights in Figure 3 reveals that the two clusters of low-scoring subsamples differed primarily in the salience of Dimension I. This difference occurred because the proportion of omitted and not reached responses was substantially greater for the cluster of low-scoring Arabic, Greek, Japanese, and Spanish than for the other low-scoring cluster or for the cluster of medium- and high-scoring examinees.

See Figure 3

Conclusions

Item Difficulty

The failure of the difficult items to contribute to defining the dimensions was unexpected. Because examinees make more errors on difficult items, these items might be expected to be more likely to cluster in ways that depend on errors. One conjecture is that difficult TOEFL items are not univocal because they involve a broad knowledge base, several distinct kinds of processes, or higher-level organizational or strategic skills that apply across many situations.

Native Language and English Proficiency

The present findings bear on the question of how many factors are measured by the TOEFL (e.g., Hosley & Meredith, 1979), and whether competence in a second language is unitary or multidimensional -- and, if the latter, what is the nature and relative importance of the various dimensions (e.g., see the review by Vollmer & Sang, 1983). The study suggests that the proficiency level of the sample exerts considerable influence on the test structure that is observed.

Implications for the TOEFL's Validity and Use

The findings have implications for the TOEFL's validity and use. The parallels between the dimensions and the sections of the test support its construct validity. The similarity in the dimensions for the different language groups suggests that the test is measuring the same constructs in each group. And the greater salience of the dimensions for the low-scoring examinees implies that the test is measuring more differentiated and distinctive constructs for these individuals.

This last finding also suggests that the interpretation of TOEFL section scores depends on the examinees' overall level of proficiency. The section scores are likely to be most useful for low scorers, helping to pinpoint the strengths and weaknesses of these individuals. In contrast, the total score is probably most useful for high-scoring examinees, providing global information about their proficiency. Follow-up research is essential to confirm the need for differential score interpretations of this kind.

The results also imply that easy and difficult TOEFL items differ in their ability to measure specific language skills and general language proficiency. The easy items appear to be the best measures of specific language skills and hence may be most useful for diagnostic purposes; the difficult items seem to be the best measures of general proficiency and thus may be most useful for global screening. This outcome raises some interesting possibilities. One possibility would be to obtain additional scores for the present TOEFL: scores based on easy items for diagnosis, and scores based on difficult items for global screening. Another possibility would be to alter what the TOEFL measures simply by changing the difficulty of the items in the test, either enhancing its diagnostic use by employing easy items or strengthening its use as a global measure by employing difficult items. Further work to understand and exploit the distinction between easy and difficult TOEFL items is clearly in order.

References

Educational Testing Service. (1985). *TOEFL test and score manual*. Princeton, NJ: Author.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-764.

Hosley, D., & Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. *TESOL Quarterly, 13*, 209-217.

Jacobson, P. E., Jr. (1976). *Introduction to statistical measures for the social and behavioral sciences*. Hinsdale, IL: Dryden.

Pruzansky, S. (1975). *How to use SINDSCAL--A computer program for individual differences in multidimensional scaling*. Unpublished report, Bell Telephone Laboratories.

Vollmer, H. J., & Sang. F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 29-79). Rowley, MA: Newbury House.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236-244.

Author Notes

The study was supported by the Test of English as a Foreign Language Policy Council.

Thanks are due Thomas S. Barrows for suggesting this study, Paul Grubb and Seymour Rosenberg for consulting on all stages of this project, and Thomas J. Jirele for computer programming.

Figure 1

Item Clusters Plotted on SINDSCAL Dimensions

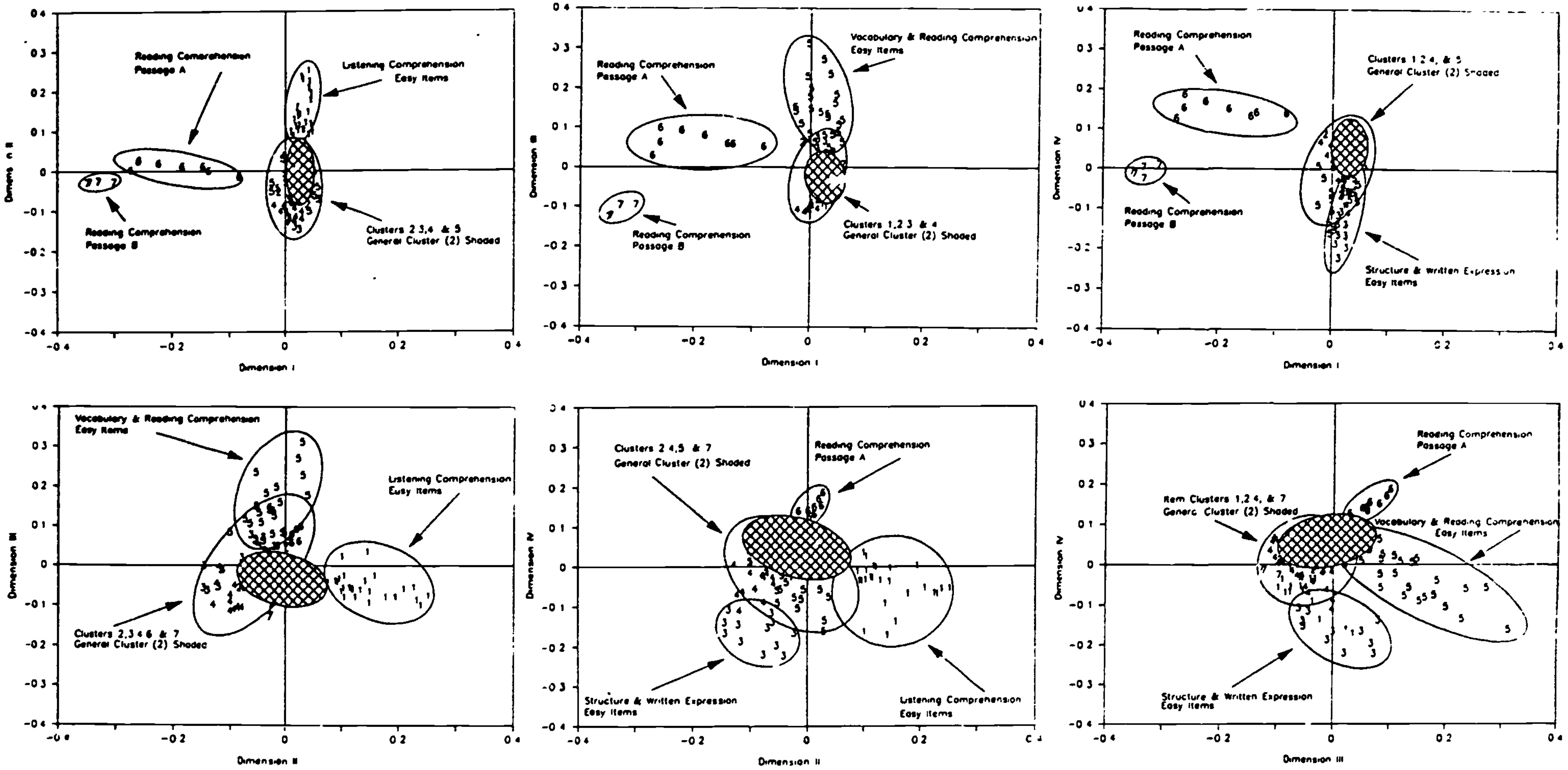
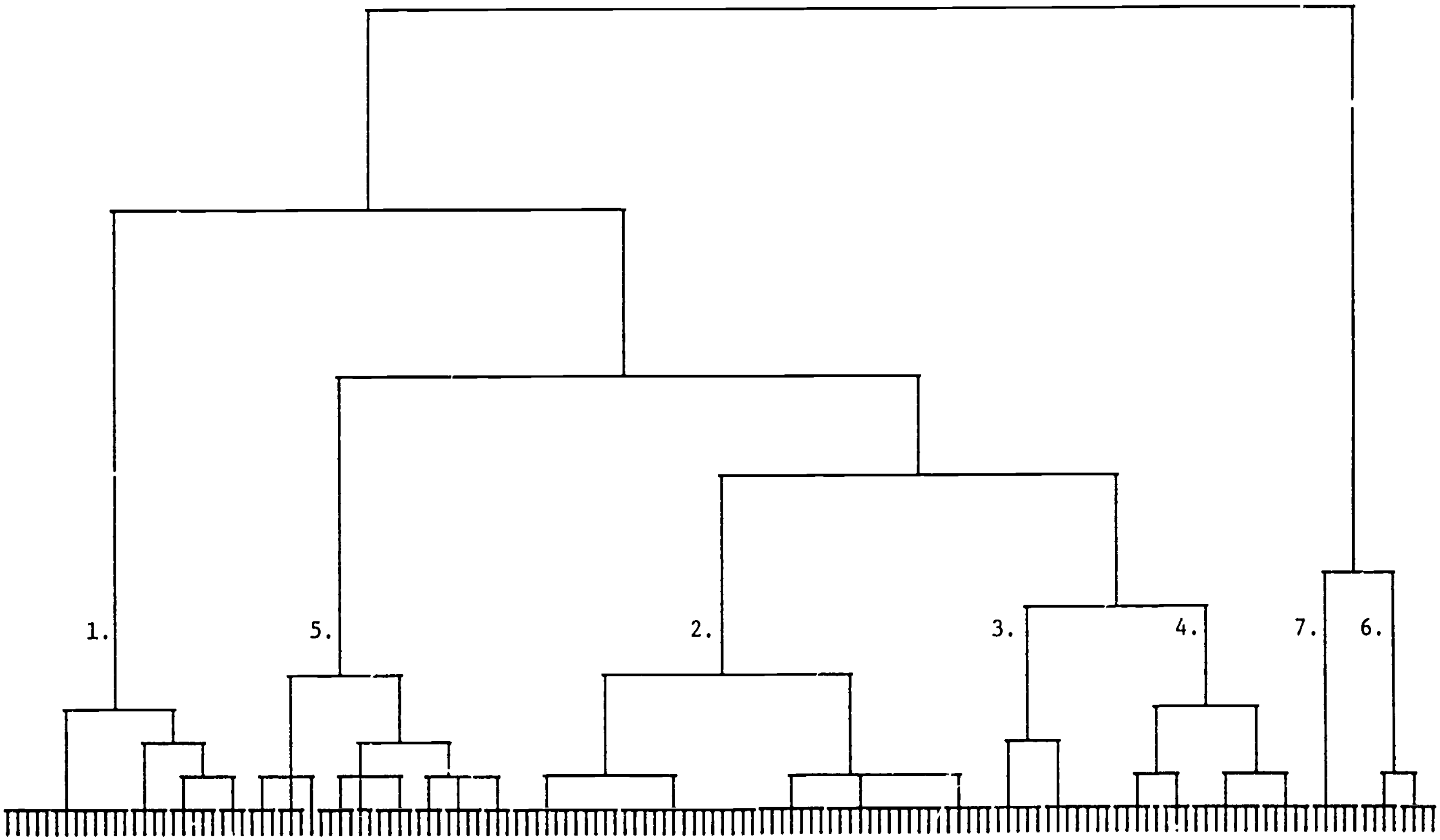


Figure 2

Hierarchical Clustering of Items

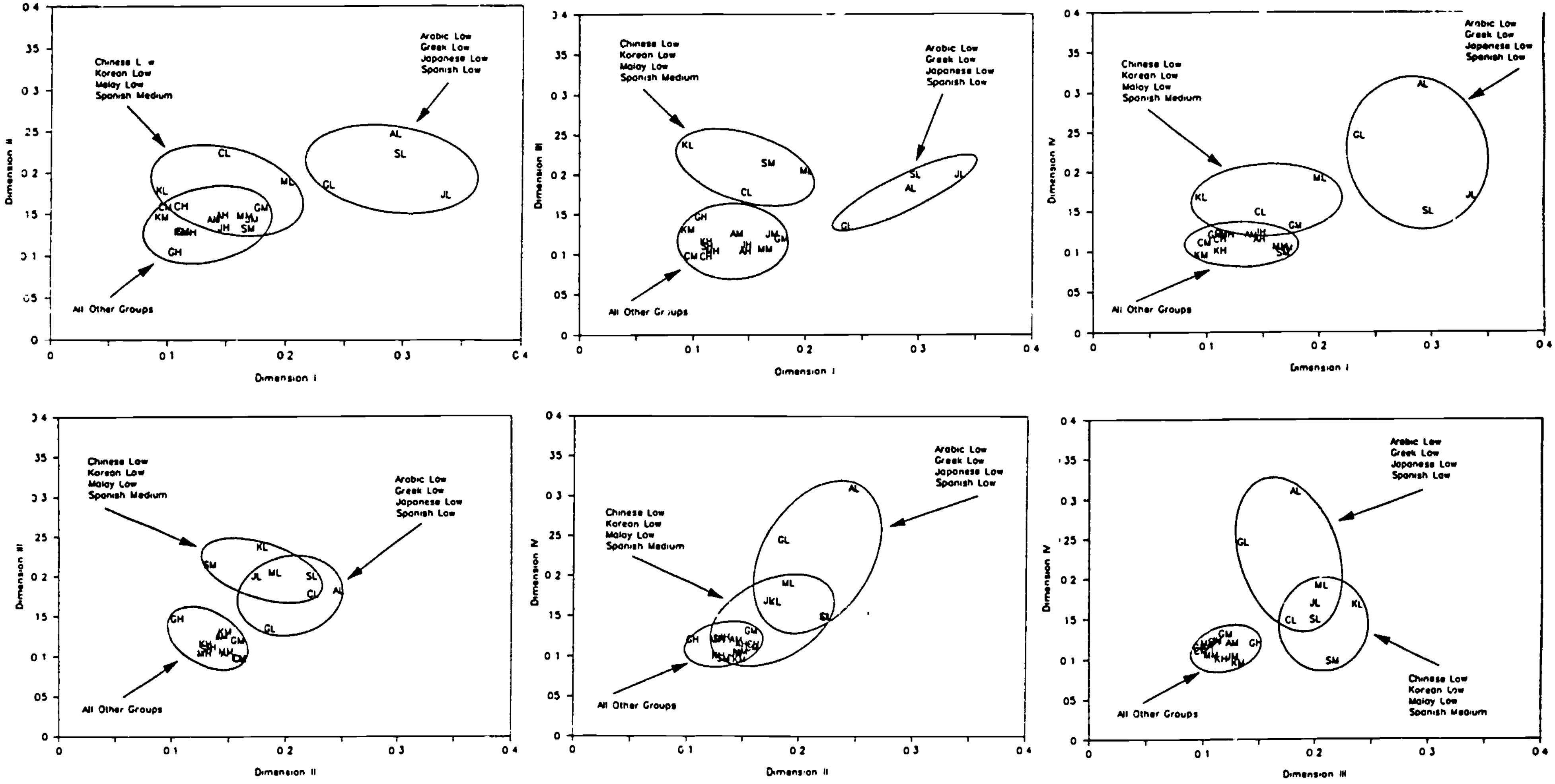


Note. Cluster composition:

1. Listening Comprehension, easy items
2. General Cluster, difficult items
3. Structure and Written Expression, easy items
4. Structure and Written Expression, medium difficulty items
5. Vocabulary and Reading Comprehension, easy items
6. Reading Comprehension, next to last passage, difficult items
7. Reading Comprehension, last passage, difficult items

Figure 3

Subsample Clusters Plotted on SINDSCAL Dimensions


Chinese Low
Korean Low
Malay Low
Spanish Medium
Arabic Low
Greek Low
Japanese Low
Spanish Low
All Other Groups
AL
SL
GL
JL
CL
KL
ML
SM
GM
CM
CH
KM
AM
AH
MM
JM
JH
GH
KH
MH
SH
Dimension I
Dimension II
Dimension III
Dimension IV