DOCUMENT RESUME

ED 287 286 FL 016 942

AUTHOR Hozayin, Russanne
TITLE The Graphic Representation of Language Competence: Mapping EFL Proficiency Using a Multidimensional Scaling Technique.
PUB DATE 87
NOTE 22p.; In: Language Testing Research; Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986); see FL 016 938.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adult Learning; Arabs; Cognitive Processes; *English (Second Language); Foreign Countries; Item Analysis; *Language Proficiency; *Language Tests; Public Administration Education; *Reading Tests; Second Language Instruction; Statistical Analysis; Test Interpretation; *Test Items
IDENTIFIERS *American University in Cairo (Egypt); *Multidimensional Models

ABSTRACT

A study investigated the use of multidimensional scaling (MDS), a statistical technique, to examine the underlying structure of an ability test at the item level. Subjects were over 8,500 Egyptian students in the Department of Public Service at the American University in Cairo. The testing instrument used was a 36-item English language test based on an English textbook used in the program. Results suggest that the cloze items, as contrasted with the elision items, are most representative of the total test score, which could be due to the relative familiarity of the item formats or to the formats themselves. Additional explanations for the results are examined. The complexity of the MDS procedure is necessary to reflect the intricacies of the cognitive processes that item-level analysis can reveal. The major drawback, however, is the present inability of specialists in the field to interpret multidimensional findings more meaningfully. Further research is recommended. (MSE)

THE GRAPHIC REPRESENTATION OF LANGUAGE COMPETENCE:

MAPPING EFL PROFICIENCY USING A MULTIDIMENSIONAL SCALING TECHNIQUE

Russanne Hozayin
American University in Cairo
English Language Institute

In a recent review of the literature on assessing the dimensionality of tests and items, Hattie (1985) pointed out that, "one of the most critical and basic assumptions of measurement theory is that a set of items forming an instrument all measure just one thing in common" (p. 139). He further noted that the assumption of unidimensionality underlies most measurement models. It also plays a crucial role in the interpretation of test results.

Two areas which have vital links with the dimensions underlying a set of items are **construct validity** and **latent structure analysis**. Traditionally, dimensionality has been closely related to the validity of the test, especially to construct validity, since investigators have often tried to assess the goodness-of-fit of a set of data to a theory-based model. That is, one approach to answering the question, "What is the test actually testing?", is to ask, "How many significant, identifiably separate factors is the test testing?"

In language ability testing, determination of underlying structure has virtually always taken the form of factor analysis of an intercorrelational matrix of subtest and/or total test scores. These subtests/test scores have most often represented a language skill (vocabulary); a specific format (cloze); or a specific test (TOEFL). (See Part I in both Oller & Perkins [1980] and Oller [1983] for detailed discussions of the pros and cons of these procedures.)

Besides construct validity, another area linked to dimensionality is that of data analysis models. In recent years, there has been a dramatic increase in the use of latent structure models[1] to analyze a wide variety of data (cf. Bergan & Stone, 1985; Young & Tanner, 1984).

Virtually all measurement models (as well as theory-based models) assume unidimensionality of test items. This assumption is particularly important for latent trait/item response theory models, like Rasch, which are

---

The author would like to gratefully acknowledge the assistance of the following people in the preparation and presentation of this paper: the faculty, staff and students of the English Language Section of the Department of Public Service, American University in Cairo; the staff of the Computer Center of AUC; and the Research and Conference Grant Program of AUC.

[1] Latent structure models are, according to Young & Tanner (1984), of four main types: latent trait, latent class, mixture analysis, and factor analysis. For one cogent approach to classification of structural models, see their article.

becoming more and more widely used in the analysis of language test results (Henning, 1984).

Therefore, in terms of the determination of construct validity of the test items as well as the fit of the measurement model to the data, the question of dimensionality is one of vital importance, both theoretically and practically. However, as Hattie (1985) has also pointed out, "despite its importance, there is not an accepted and effective index of the unidimensionality of a set of items" (p.139).

One major reason for the lack of progress in determining dimensionality of tests is that of confusion over definition. Although a clear, acceptable definition of unidimensionality is not yet forthcoming, certain clarifications have been made recently. For example, Hattie (1985) has drawn distinctions between "homogeneity," "internal consistency" and "reliability" on the one hand and "unidimensionality" on the other.

In his extensive review of more than 90 articles, Hattie made the very important point that high reliability (i.e., alpha and its derivatives) by itself may not be an indication of unidimensionality. This conclusion is crucial in view of the importance of reliability as an index of unidimensionality--"Perhaps the most widely used index of unidimensionality has been coefficient alpha (or KR-20)" (Hattie, 1985, p.143).

Henning, Hudson and Turner (1985) provided further insight when they noted that "no test can strictly be said to measure one and only one trait" (p.142). Thus, a test will be unidimensional when it "measures a primary or dominant factor" (Henning, et al., 1985, p.145) or when "one latent trait underlies the data" (Hattie, 1985, p.139). In effect, the question has been rephrased from the more simplistic "What is unidimensionality?" to the more complex, yet potentially more realistic, "How close is a set of items to being a unidimensional set?" (Hattie, 1985 p.159) and "To what extent must [unidimensionality] be present in order for latent trait models to apply?" (c.f., Henning, et al., 1985, p.142).

Thus, attempts at clarifying definitions have led to some refinements of one of the questions to which answers are being sought. However, beyond problems with definitions there are difficulties with the statistical techniques which have been applied to the problem of unidimensionality, particularly with factor analysis.

In the past, there have been two main approaches to the assessment of dimensionality of tests or items:

(1) The investigator has a set of data which he or she wishes to analyze in order to reveal the number of factors being measured.

data ---> via measurement method ---> model

(2) The investigator has posited a model and he or she wishes to assess the goodness-of-fit of that model to a set of data.

model ---> via data analysis ---> fit

For both of these approaches, in language testing, subtest scores or scores from a variety of instruments have been correlated, and the resulting matrix has been factor analyzed--by using either traditional factor analytic methods (Oller & Hinofotis, 1980) or confirmatory factor analysis (Vollmer, 1985). In order to be able to assess the relative success of these past approaches, both the growing importance of item level analysis and the increasing criticism of factor analysis on both technical and substantive grounds must be taken into consideration.

As has already been noted, the increasing interest in item-level analysis in language testing has been due in large part to the increased use of latent trait analyses, which have been stimulated by the rapid developments in computer technology as well as the general spread of cognitivism. In addition, with such developments as item banking and computer adaptive testing, wherein relatively few items can give the same level of information on the subject's ability as can a longer paper-and-pencil test, item-level analysis is becoming more and more important. Besides the importance of item-level analysis, there are two critical points about factor analysis that are pertinent to the current study, including assumptions about the linearity of relationships among the factors as well as requirements for certain types or levels of data when using different latent structure techniques.

First, factor analysis--as a latent structure technique--is intended to reveal the "hidden" relationship among the variables and thereby lay bare the underlying structure of the trait (items) being measured. However, in nearly all previous research studies which used factor analysis, it was assumed that the relationships between the scores used to generate the correlation matrix were linear. This assumption must be looked at more closely, in the light of recent developments in measurement models and the assumptions underlying them.

As Vollmer & Sang (1983) pointed out,

> . . . possibly a linear outlook on language proficiency is far too simple...it is important to stress that . . . differences in the interaction between different levels and aspects of ability on the one hand and between object (items) and subject (person) on the other hand can neither be discovered nor be described by means of classic factor analysis (p.72).

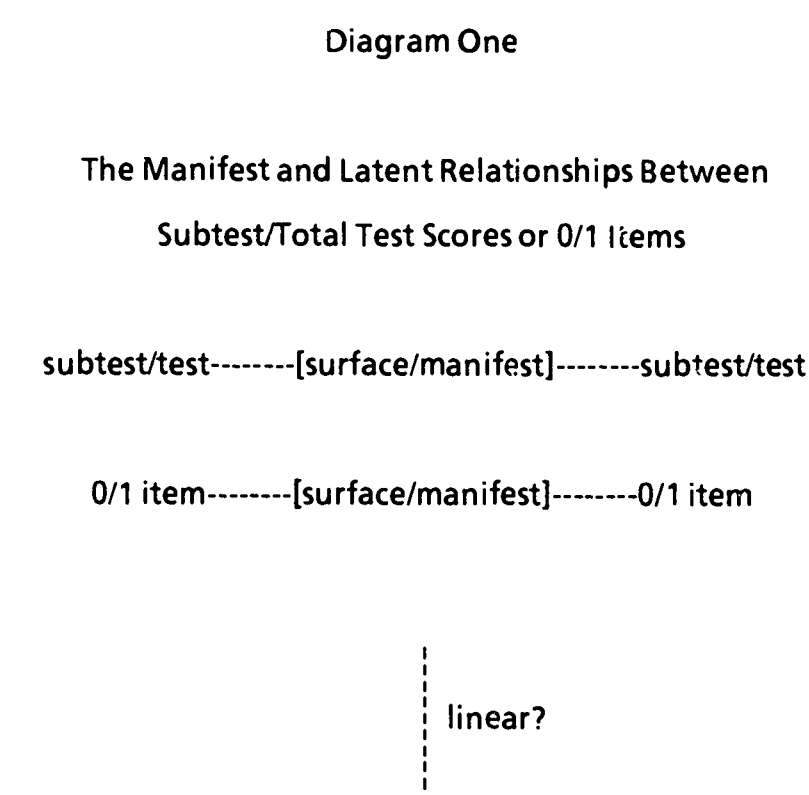
Other writers concur, as when Hattie (1985) stated that, "a further problem of many procedures is that a linear model cannot be assumed. When items are scored dichotomously, then the use of a linear factor model [is] not appropriate since [it] assumes linearly related variables" (p. 158).

It must be recalled here that there are in fact at least two relationships whose linearity/non-linearity must be accounted for. Diagram One represents the relationship between the manifest scores (or item responses), on the one hand, and between the surface representative and the latent trait which the test (item) score is assumed to reflect, on the other. The top line depicts the manifest relationship between two subtest/total test scores (traditionally calculated by a Pearson product moment correlation coefficient); the second line, the manifest relationship between two 0/1 items (traditionally calculated by a phi or tetrachoric coefficient).

Diagram One

The Manifest and Latent Relationships Between

Subtest/Total Test Scores or 0/1 Items



The vertical line in Diagram One represents the relationship between the manifest score, whether subtest/total test or 0/1 item, and the posited latent trait. This relationship is based on a theoretical model and its nature at present must be considered to be largely speculative, due to the paucity of results based on analysis of actual data sets.

The second point related to the use of factor analysis is that of the level of the data. One of the crucial factors in the choice of a statistical technique for a given data analysis situation is whether the data is categorical, rank-order, continuous or ratio. Young and Tanner (1984) have pointed out that factor analysis is appropriate when both the manifest and latent variables are continuous, while latent trait techniques are suitable when the manifest variable is dichotomous and the latent is continuous. Therefore, at the item level, if the items are dichotomous, factor analysis is inappropriate. It may be concluded, then, that if item level test analysis is to be undertaken, then linear factor analysis is in fact an ineligible analytic technique, both due to the questions raised above concerning the nature of the manifest-latent relationships and to the level of the data involved in 0/1 item analysis.

As an alternative to linear factor analysis--which was formulated in a time when technology was relatively embryonic and whose assumptions may well be unrealistic in terms of reflecting the nature of the relationships among the structures underlying person responses, it was decided to try another

approach to assessing unidimensionality--that of non-metric multidimensional scaling, using item-level analysis.

Multidimensional Scaling

Multidimensional scaling (MDS) is a statistical technique often compared to factor analysis (FA):

> they are both used to study the structure of objects; they require similar input data, proximity measurements defined over pairs of objects; [and] they both represent structure in terms of spatial coordinates (Davison, 1985, p. 94).

As Schiffman, Reynolds and Young (1981) point out, however, there are certain crucial differences between MDS and factor analysis, particularly that "the MDS model is based on distances between points whereas the FA model is based on angles between vectors...[the former being] easier to interpret" (p. 13). In addition, they point out that "FA often results in a relatively large number of dimensions mainly because most procedures are based on the assumption of linear relationships between the variables ..[MDS] does not contain this assumption" (p. 13).

An important advantage of MDS over factor analysis is the graphic representation of the data which comprises part of the MDS output. To obtain this pictorial analysis, MDS programs present numerical representatives of objects which are similar to one another (in this study, correlations between 0/1 responses to items) as points close to each other in a spatial map. Objects which are dissimilar are represented as points distant from one another (Schiffman, et al., 1981). Since MDS uses distance instead of angular separation (which is used by factor analysis), there is no assumption of linearity required (Coxon, 1982).

Subkoviak (1975), in reviewing the use of MDS in educational research, confirmed that MDS "can be used to discover the underlying structure of a test" (p. 412) at the item level, although it has rarely been used for this purpose. More recently, Davison (1985) noted that, "Factor analysis... remains the most common method used to analyze the structure of tests or test items. More and more, however, MDS has been recommended for the same purpose"(p. 94). Those who have used it at the item level include Farley & Cohen (1974), Karni & Levin (1972), and Napior (1972).

Prior to this study, as far as can be determined, MDS has never been used to investigate the structure underlying an ability test at the item level as an alternative to factor analysis. After reviewing the literature, though, the present investigator concluded that because MDS does not require an assumption of linearity, and because its output is in the form of graphical representations of the data, it would be a viable alternative to factor analysis.

However, it must be assumed for the purposes of this study that the relationship between answers to each pair of items for a given sample of subjects can be analyzed to reveal information about the structure of the language competence of those subjects. With this assumption in mind, the methods used in the study, along with the results and conclusions, will now be presented.

Method

Subjects

The subjects who participated in this study were part of a group of more than 8500 students who take adult English courses at the Department of Public Service (DPS), American University in Cairo, each term. These courses are non-intensive (3 hours per week), and last for 12 weeks In the majority of cases, students do not take these courses in order to satisfy an immediate job or educational requirement, but in order to improve their English skills, possibly to obtain a better position in the future.

The subjects in the current study were 149 adult Egyptian basic-level EFL learners, aged between 18 and 55. These people were placed in Level 3 (out of 17 levels) on the basis of: (1) their scores on a standardized test written and used by the DPS for placement, or (2) their having passed Level 2. Following their placement at Level 3, they were then assigned at random to one of the 36 Level 3 classes which were offered during the Fall, 1985 term. Students in all 36 classes took the post-test which formed the basis of this study. Class-size averaged between 14 and 18 students.

Instrument

The test used in this study was one of three forms taken by the Level 3 students as a post-test. These three forms were the result of a pre-testing process which took place over a 9 month period and included approximately 2500 students from Levels 2 through 5 of DPS. Each form contained 36 questions, including six multiple-choice grammar and vocabulary items, two 10-item cloze passages and one 10-item elision passage. In the latter, the students were required to delete 10 extraneous words from the passage. The ten target words, however, were not indicated in any way, so the students had a discourse level rather than a discrete-point type of task (although, due to the low level of their English ability, they were assisted in completing the task by being told the precise number of extra words which occurred in the passage).

All items were based on the Key to English textbooks used in the DPS courses. These books were designed and written by the DPS curriculum developers specifically for Egyptian students. Prior to the item writing process, a detailed content analysis of the textbooks was carried out (Hozayin, 1986), with the aim being to establish a basis for determining the content validity of the test.

After each item pre-testing session, the results were item-analyzed, using the Rasch BICAL program (Wright & Stone, 1979). Items for subsequent forms were selected on the basis of their total T-Fit (a BICAL statistic) being equal to or greater than one standard deviation below the mean total T-Fit. Other criteria that were taken into consideration in selecting items were the discriminability of the item (.80 or higher) and the point biserial correlation (.25 or higher). These additional criteria were used since at times the T-Fit of the item may be acceptable, but the item may not add much information concerning the students' ability, as, for example, when an item is too easy.

This pre-testing process resulted in short tests (36 to 38 items) with relatively high KR-20 reliabilities (typically, between .86 and .92 for the total test). As may be seen in Table 1, the KR-20 reliability of the total test was .89, while those of the two cloze and the one elision passage were .79, .71 and .80, respectively.

Table 1 also shows the means and standard deviations of the total test and the three passages. The mean of 18.61 (out of a possible 36), which is lower than one might expect on a post-test, probably resulted from a combination of circumstances. First, the item pre-testing process showed that items which fit the students prior to the course were not suitable post-course; thus, the pre-and post-tests were not exactly the same, although they did share some items. Second, the non-intensive nature of the program and the low level of the students, as well as their lack of immediate, specific and vital purpose in taking the course, could be acting (and inter-acting) to depress the mean.

Another factor, mirrored in the relatively high standard deviation (7.25 for the total test), is that although the students were all basic-level, they were still characterized by a range of abilities with their general level. This is because placement, as mentioned above, may have been by screening test scores or by the students having passed the previous level. This two-pronged system, when combined with student and program attributes, led to diversity of ability among the students. It also may be in part responsible for the elevated reliability of the test.

The subtest and total test intercorrelation coefficients contained in Table 1 show that there were moderate correlations between the three subtests (from r = .4294 to r = .5619). It should be mentioned here that multiple choice items 5 and 6 were included in the correlation matrix because they were the only multiple choice items identified by the multidimensional scaling analysis (see below) as being significantly related to the trait under discussion. In addition, since the six multiple choice items were analyzed as individual items, not as a subtest, means and standard deviations were not reported for these items by themselves or as a subtest.

Table 1

Test Statistics

Test Subtest	Mean	SD	DR20*	No. of Items
Total	18.61	7.25	.89	36
Cloze 7-16	4.76	2.72	.79 (.88)	10
Elision 17-26	5.80	2.43	.80 (.88)	10
Cloze 27-3	4.83	2.30	.70 (.83)	10

*Due to the small number of items per subtest, the Spearman-Brown formula was applied. The estimated r for 20 items is given in parentheses.

N = 149 for all tables and graphs.

Table 1 (Continued)

PMC Intercorrelation Matrix of Items, Subtests and Total

Item 5	1.000					
Item 6	.2617	1.000				
Cloze 1	.4813	.4407	1.000			
Elision	.3610	.2810	.5389	1.000		
Cloze 2	.3340	.3370	.5619	.4294	1.000	
Total*	.5002	.4381	.7156	.5585	.5702	1.000
	Item 5	**Item 6**	**Cloze 1**	**Elision**	**Cloze 2**	**Total**

*These correlations have been corrected for part-whole overlap.

Data Collection Procedures

The tests were administered by the classroom teachers in the last week of class. The three forms were distributed at random to the students. Afterwards, the tests were returned to the investigator who scored them according to criteria derived from the responses of twenty EFL teachers (12 native speakers of English and 8 non-native speakers) to the cloze and elision passages. The responses of the native speakers were used in the cloze correction process. In the rare cases when a potentially correct response was not included among the answer given by the native speaker teachers, two native speakers (the investigator and one teacher) decided whether to score the response as being correct or not.

The elision passage was scored by giving credit only when the ten items of interest were crossed out. If other words in the passage were crossed out, no points were deducted. Rarely did a student cross out more than ten words. If they got a low score on elision it was usually because they couldn't perceive that a word was extraneous (thus deleting fewer than 10 words), or they perceived the extra word incorrectly (deleting 10 words, but not the correct ones). Scoring procedures for elision are obviously more complex than those used with cloze. The scoring procedure applied in this study was used in order to obtain a stable data set for inclusion in the analysis. However, in the future, alternative procedures could be applied and the resulting data could be re-analyzed following the steps outlined below.

Data Analysis

The scores for the 36 items were first analyzed using BICAL. Then a Pearson inter-item correlation matrix was generated. These correlation coefficients were then analyzed by KYST, a non-metric MDS program (see Kruskal, Young,

& Seery, 1973, for a complete description of KYST). Since a non-metric MDS application was being used, Pearson coefficients[2] were acceptable input for the KYST program (cf. Coxon, 1982, for a discussion of the types of data which may be used in MDS programs).

The KYST output, which included data plots for 1 through 5 dimensions, reached a minimum configuration in all five dimensions, with the following stress rates (error or noise)[3]:

1 dim = .418; 2 dim = .267; 3 dim = .198;

4 dim = .155; 5 dim = .126

Although there was a noticeable decline in stress as the number of dimensions increased, the level of stress in five dimensions was still relatively high, according to guidelines offered by Kruskal and Wish (1978). This may have been due to: (a) the basic level of the students; (b) the fact that the analysis was at the item level; (c) the nature or number of the items; or, most likely, (d) a combination of these factors. The implications of these factors in terms of stress are discussed below.

Results and Discussion

Since correlations were used in the KYST analysis, the emphasis was on small distances (large similarities) between pairs of items. Therefore, the results are presented in the form of neighborhood analyses (cf. Kruskal & Wish, 1978). That is, as shown in Table 4, below lines were drawn between the 32 item pairs with the highest correlation coefficients, ranging from .384 through .581. These coefficients represent 5 percent of the coefficients used in the KYST analysis, which totalled 630, this being the total number of coefficients arising from correlating 36 items and creating a half-matrix (i.e., number of items divided by number of items minus 1, divided by 2 to give a half-matrix).

Appendix A contains the 22 items identified by this procedure, while the item statistics, including Rasch total T-fit, difficulty and point-biserial correlations are given in Table 2. In addition, the item characteristic curve data

---

[2] Although such a discussion is beyond the scope of this paper, there are several very important considerations in selecting correlation coefficients for use in multivariate procedures. The interested reader should consult Coxon (1982) for specific information on MDS, Harris (1975), and Thorndike (1978).

[3] While the definition of stress is fairly standard in any work on MDS, for example, "[stress is] a measure that shows how far the data depart from the model" (Schiffman, et al., 1981), the interpretation of the stress coefficient and determination of what an acceptable level of stress is in a given case are both rather vexed subjects in the literature. Kruskal and Wish (1978) offer the most fundamental explanation, which a newcomer to MDS may wish to pursue.

are given in Table 3, showing the percentage of subjects who got each item correct, for six ability groupings. These 22 items may be said to represent the core of the learner's EFL competence, as measured by this test.

Table 4, (a, b, c, d) which offers a graphic portrayal of the structural relationship of the items that make up this core, contains a neighborhood KYST analysis of 22 core post-test items for these 149 subjects, for 2 by 1 dimensions in 2-, 3-, 4- and 5-dimensional space. In order to assess the dimensionality of the data, these diagrams were first examined for patterns in the data according to four types of criteria: (1) format/item type; (2) item fit (Rasch T-Fit from BICAL); (3) item content; and (4) item difficulty.

Although not conclusive, the only clear pattern was based on clustering by item type, in that several of the elision items were grouped together on the left of the diagram, while most of the cloze items were grouped in the center (see Table 4). Thus, the cloze items appear to form the core of the diagrams, being the items which are most representative of the total test score (cf. their higher point-biserial correlations). Further, it is interesting to note that there are two types of relationships possible among the items--spatial proximity and correlational proximity. For example, the four cloze items from one of the two passages included in the test (i.e., items 33-36) formed one separate cluster (in terms of their high intercorrelations). At the same time, they were also in fairly close proximity to other cloze items, certainly closer to them than to the elision items which were on the opposite side of the diagram.

These relatively distinct groupings of elision and cloze items may indicate that two different cognitive activities are being measured by this test. It should also be kept in mind that while the students were familiar with cloze procedures, they were not as familiar with elision procedures, according to several of the teachers. This lack of familiarity may be responsible for the apparent division between these two formats. In subsequent test forms, instructions were given in Arabic and teachers reported that difficulties associated with elision were much reduced. The results of these tests are being examined to determine if patterns similar to those found in this data set are typical or if these patterns are related to the subjects' unfamiliarity with the elision format.

A second approach to evaluation of the diagrams was to examine the 2 by 1 dimensional structure underlying the 22 core items. When each of the four dimensions (Table 4) was assessed, it was noted that the five-dimensional portrayal of the relationship between dimensions one and two was the most parsimonious, since the items which were most closely related (i.e., which had the highest correlations) were also closest together in space in the five dimensional diagram. This is a further indication that these items are measuring multiple objects.

What conclusions may be drawn from this application of multidimensional scaling to an item-level EFL ability data set? With regard to the issue of unidimensionality, in terms of both construct validity and models for latent structure analysis, several questions may be raised, chief among them being whether it is possible that items might be multidimensional (as they appear to be here), while the total test could be considered to be unidimensional in the currently accepted definition of the term. However, even if this were found to be true, it is clear that if item level analysis is to be pursued, even the recent,

more explicative definitions of unidimensionality mentioned in the Introduction will need to be reconsidered.

Equally clearly, much more use needs to be made of MDS at the item level. It should be mentioned here that this application is exploratory in nature rather than being confirmatory MDS (cf. Traub & Lam, 1985; Young, 1984). In future studies, if insights obtained into item-level analysis of ability by similar applications of exploratory MDS warrant, then confirmatory MDS might be used.

Another consideration is that this particular set of data has a relatively high amount of stress. Possible explanations already offered are related to the ability level of the students, the item-level of the analysis, the number and type of items or a combination of these factors. It might be noted in this connection that a reanalysis by the present author of the Oller-Hinofotis (1980) [following Wang's (1984) example] total test/subtest data using KYST revealed a much lower stress value (.01 in five dimensions) than was found for the current item-level data. Thus, it may be that item-level data have more noise, or that this particular data set has more noise.

Further analyses at the item level are needed in order to shed more light on the problem of stress or noise in the data. Future studies could focus on item analysis of longer tests as well as of standardized tests, using results obtained from EFL students of differing abilities.

It is readily apparent that that MDS is a complex procedure. However, the current investigator believes that this complexity is necessary in order to be able to reflect the intricacies of the cognitive processes which, it is assumed, item-level analyses of ability test results can reveal. Somewhat offsetting the possible disadvantages of this complexity is the spatial output one may obtain, which may be analyzed to reveal information concerning the product of language learning and--after much more use is made of it--possibly the process.

The major drawback to multidimensional scaling at this point is our lack of ability to interpret the multidimensional findings more meaningfully, especially in terms of the implications they may have for the construct validity of language ability tests, further explication of which was one of the purposes stated at the outset of this paper. Capitalizing on the spatial nature of the MDS output is a potentially fruitful path to meaningful interpretation of findings, since several other disciplines (such as sociology, cognitive psychology, ecology, biology, and physics) have established guidelines for analysis of spatial data. The current investigator is pursuing several analytic methods used in other fields in an attempt to shed light on the findings reported herein.

With regard to the assumption of unidimensionality underlying latent trait or latent structure analyses, it may be noted that investigations into dimensionality as they now stand include a large element of informed intuition and judgement on the part of the investigator, which may also be characterized as subjectivity. A certain amount of this seems to be unavoidable given our present state of knowledge (i.e., the lack thereof).

When a complex subject is being explored, summary methods of analysis -- although perhaps more parsimonious -- may not produce results which can adequately reflect the nature of the object of study. In order for investigators to begin to draw nearer to the desiderata of parsimony and elegance in theory

building, model testing, and data analysis, approaches which are systematic yet which produce richer, more explicative and potentially more meaningful results are necessary. It is hoped that the application of MDS to item level ability data described in this paper will prove to be a step in this direction.

References

Bergan, J. & Stone, C. 1985. Latent class models for knowledge domains. Psychological Bulletin, 98(1):166-184.

Coxon, A. 1982. The User's Guide to Multidimensional Scaling. New York: Heinemann Educational Books.

Davison, M. 1985. Multidimensional scaling versus components analysis of test intercorrelations. Psychological Bulletin, 97: 94-105.

Farley, F. & Cohen, A. 1974. Common-item effects and the smallest space analysis of structure. Psychological Bulletin, 81: 766-772.

Harris, R. 1975. A Primer of Multivariate Statistics. New York: Academic Press.

Hattie, J. 1985. Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9 (2):139-164.

Henning, G. 1984. Advantages of latent trait measurement in language testing. Language Testing, 1:123-133.

Henning, G., Hudson, T. & Turner, J. 1985. Item response theory and the assumption of unidimensionality for language tests. Language Testing, 2:141-154.

Hozayin, R. 1986. Using curricular domain assessment to establish the content validity of an EFL achievement test. A paper presented at the 6th National Symposium on English Teaching in Egypt: Testing and Evaluation, Ismalia, Egypt, March 25, 1986.

Karni, E. & Levin, J. 1972. The use of smallest space analysis in studying scale structure: An application to the California Psychological Inventory. Journal of Applied Psychology, 56:341-346.

Kruskal, J.B. & Wish, M. 1978. Multidimensional Scaling. Beverly Hills, CA: Sage Publications.

Kruskal, J.B., Young, F.W. & Seery, J.B.1973. How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Murray Hill, NJ: Bell Labs., mimeo.

Napior, D. 1972. Nonmetric multidimensional techniques for summated ratings. In R. Shepard, et al., Multidimensional Scaling, Vol. 1. (New York: Seminar Press).

Oller, J., ed. 1983. Issues in Language Testing Research, Rowley, MA: Newbury House .

Oller, J. & Hinofotis, F. 1980. Two mutually exclusive hypotheses about second language ability: indivisible and partially divisible competence, in Oller & Perkins, 1980.

Oller, J. & Perkins, K. 1980. Research in Language Testing, Rowley, MA: Newbury House.

Pang, L.Y. 1984. Is there a global factor of language proficiency? A critique of Oller and Hinofotis (1980). IRAL, 22:203-212

Schiffman, S., Reynolds, M. & Young, F. 1981. Introduction to Multidimensional Scaling, New York: Academic Press.

Subkoviak, M. 1975. The use of multidimensional scaling in educational research. Review of Educational Research, 45:387-423.

Thorndike, R. 1978. Correlational Procedures for Research, New York: Gardner Press (distributed by Halstead Press).

Traub, R.E. & Lam, Y.R. 1985. Latent structure and item sampling models for testing. Annual Review of Psychology, 36:19-48.

Vollmer, H. 1985. Models of second language competence: A structural equation approach. A paper presented at the 7th Annual Language Testing Research Colloquium, Princeton, N.J., April 6-9, 1985.

Vollmer, H. & Sang, F. 1983. Competing hypotheses about second language ability: A plea for caution, in J. Oller (1983), pp. 29-79.

Wright, B. & Stone, M. 1979. Best Test Design, Chicago: Mesa Press.

Young, F.W. 1984. Scaling. Annual Review of Psychology, 35:55-81.

Young, M.A. & Tanner, M.A. 1984. Recent advances in the analysis of qualititative data with applications to diagnostic classification. In Gibbons, R.D. & Dysken, M., eds. Statistical and Methodological Advances in Psychiatric Research. New York: Spectrum.

Appendix A

Items Included in Graphs

Items (M - Multiple choice vocabulary; C - Cloze; E - Elision)

M5 - I can't hear you because the boys are very _____.
1. noisy 2. busy 3. tired 4. important

M6 - Please find this word in the _____.
1. calculator 2. dictionary 3. typewriter 4. adding machine

C7-16 (Items 8 and 15 not included in graphs):

Mona has a big family. __C7__ has three brothers and three __C8__. Two of her brothers __C9__ engineers and the other is a student. __C10__ of her sisters is married, but the __C11__ are studying business at the university. Mona's __C12__ is an accountant and __C13__ mother is a teacher __C14__ a girl's school. Because __C15__ are many people in the family, __C16__ must work hard to succeed.

E17-26 (Items 25 and 26 not included in graphs):

"Hello, John. How are you?" "Fine, thanks, Ahmed. How are you?"
"Fine, thanks. I'm going [E17 at] to the Pyramids. Can you come with [E18 to] me?"
"No, I'm sorry. I [E19 am] can't. I'm busy?"
"That's too bad. Next week I'm going [E20 go] to visit my sister in [E21 the] Ismalia. Can you go with [E22 my] me?"
"OK. What time are you going [E23 at] to leave?"
"I am leaving [E24 in] at 8 a.m. Where can I pick [E25 get] you up?"
"In front of my house [E26 family]."
"OK. See you next week, John."

C27-36 (Items 27 through 32 not included in graphs):

"What's the matter, Nabil? Are __C27__ tired?"
"No, I'm not. I'm __C28__. Is there a restaurant __C29__ here?"
"Yes, there's a good one __C30__ there."
"What's the food __C31__?"
"It's very good. What do you __C32__ to eat?"
"I'm not sure. __C33__ there any kofta?"
"Yes, there's __C34__, and there's kebab too."
"Are __C35__ any good salads?"
"Yes, there __C36__ some."
"OK. Let's go eat."

Table 2

Rasch Statistics

Item	T-Fit	Diffic	pbs
M5	-.70	.06	.55
M6	-.28	.06	.51
C7	-.58	-1.42	.43
C9	-.83	.49	.55
C10	-1.56	.26	.63
C11	-1.51	1.36	.59
C12	-1.93	.51	.66
C13	-1.47	-.39	.61
C14	-1.04	-.24	.57
C16	-1.27	.31	.60
E17	-.50	-1.18	.44
E18	-1.29	-1.03	.55
E19	-.68	-.94	.50
E20	-.27	-.56	.48
E21	-.43	.96	.48
E22	-.79	-1.42	.48
E23	-.91	-1.18	.48
E24	-1.06	-.76	.55
C33	-.94	-.37	.56
C34	-.85	-.42	.54
C35	-.69	.03	.55
C36	.12	-.09	.47

T-Fit = Rasch item total t-fit Diffic = Rasch item difficulty
pbs = point biserial correlations

Table 3

Item Characteristic Curve

Item	Lowest	2	3	4	5	Highest
M5	.16	.19	.50	.65	.77	1.00
M6	.28	.31	.39	.54	.92	1.00
C7	.60	.73	.89	.96	1.00	1.00
C9	.08	.12	.32	.46	.69	.92
C10	.00	.23	.32	.65	.77	1.00
C11	.00	.00	.04	.02	.58	.77
C12	.08	.00	.29	.38	.88	.92
C13	.12	.50	.57	.85	.96	.92
C14	.16	.31	.57	.88	.81	1.00
C16	.08	.15	.29	.62	.81	.92
E17	.40	.73	.93	.92	1.00	.92
E18	.32	.62	.93	.96	.96	1.00

Table 3

'.em Characteristic Curve (Continued)

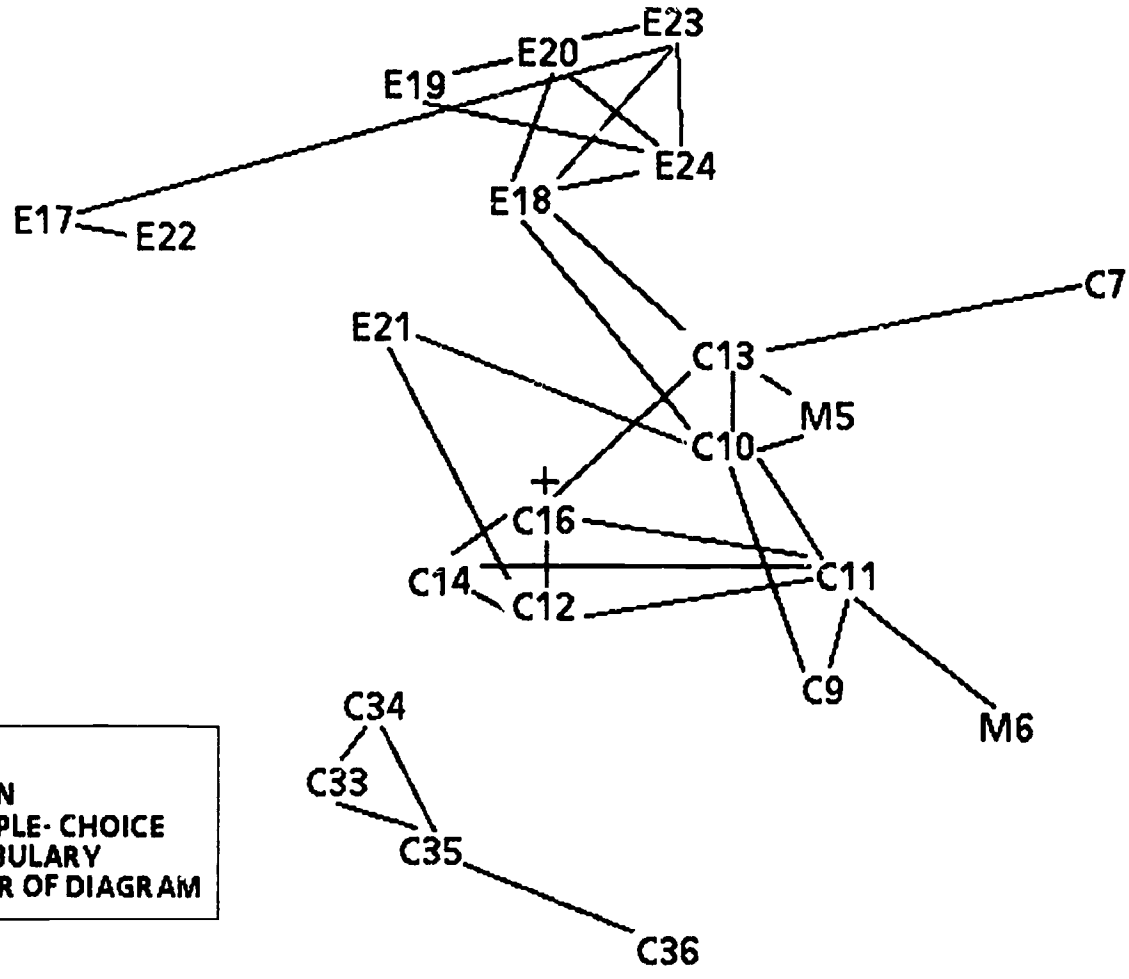
Item	Lowest	2	3	4	5	Highest
E19	.36	.65	.75	.92	1.00	1.00
E20	.32	.46	.68	.85	.92	.92
E21	.04	.08	.14	.38	.50	.96
E22	.52	.73	.93	1.00	1.00	1.00
E23	.36	.81	.89	.92	.96	1.00
E24	.32	.54	.71	.96	.92	1.00
C33	.20	.35	.64	.88	.85	1.00
C34	.20	.46	.61	.92	.81	1.00
C35	.16	.31	.39	.58	.88	1.00
C36	.28	.42	.32	.62	.88	1.00
Score range:	1-10	11-14	15-19	20-24	25-28	29-35
No. of SS:	25	26	28	26	26	13

Note: Figures represent percentages in each ability grouping

Table 4a

TWO DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

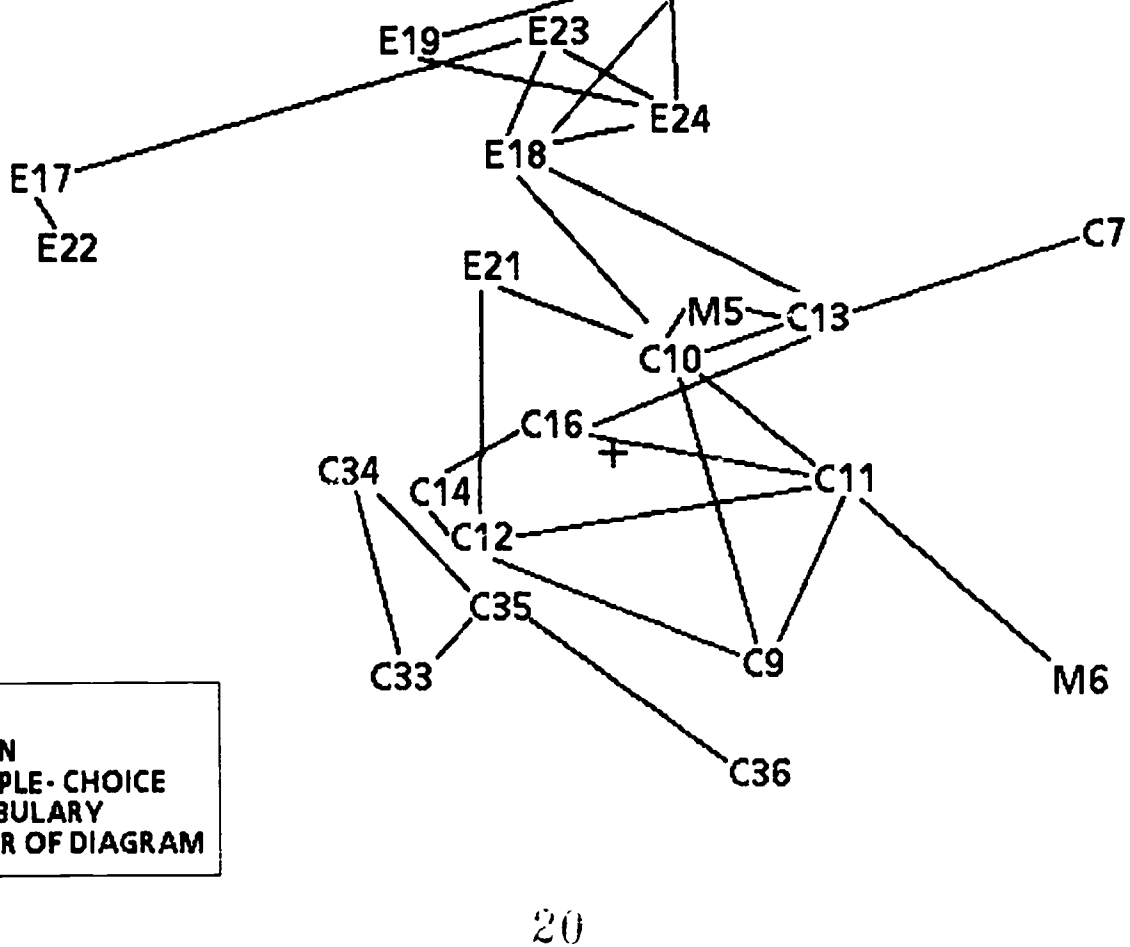


C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

Table 4b

THREE DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

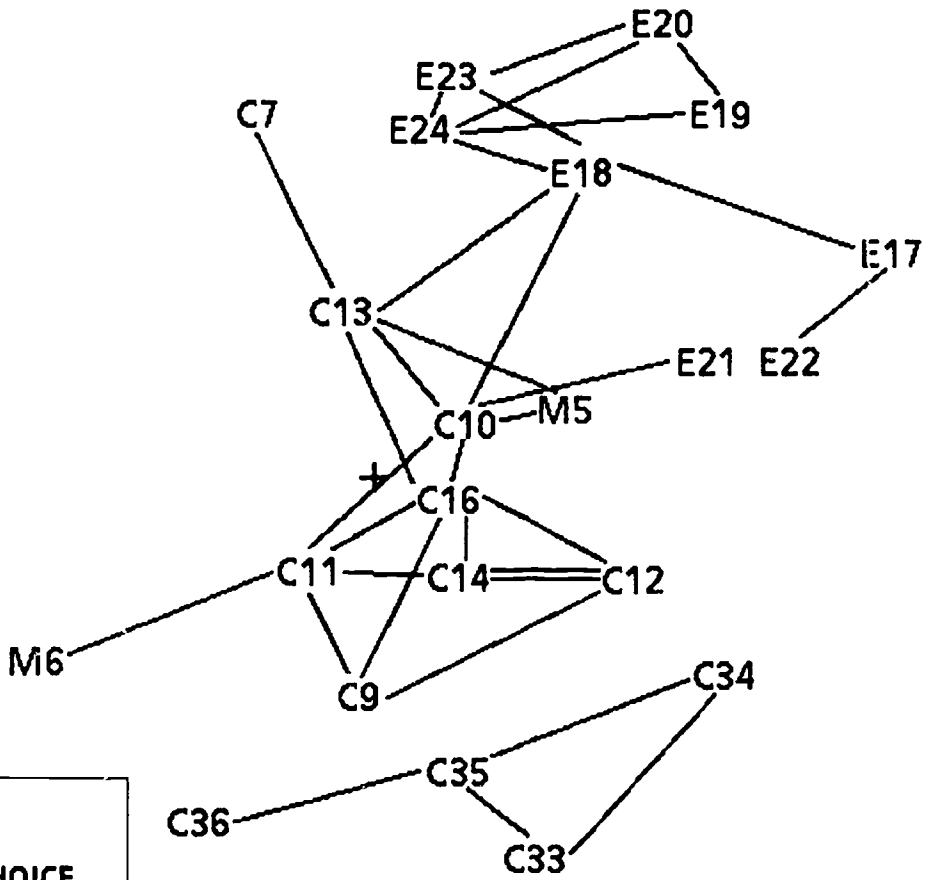


C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

Table 4c

FOUR DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS

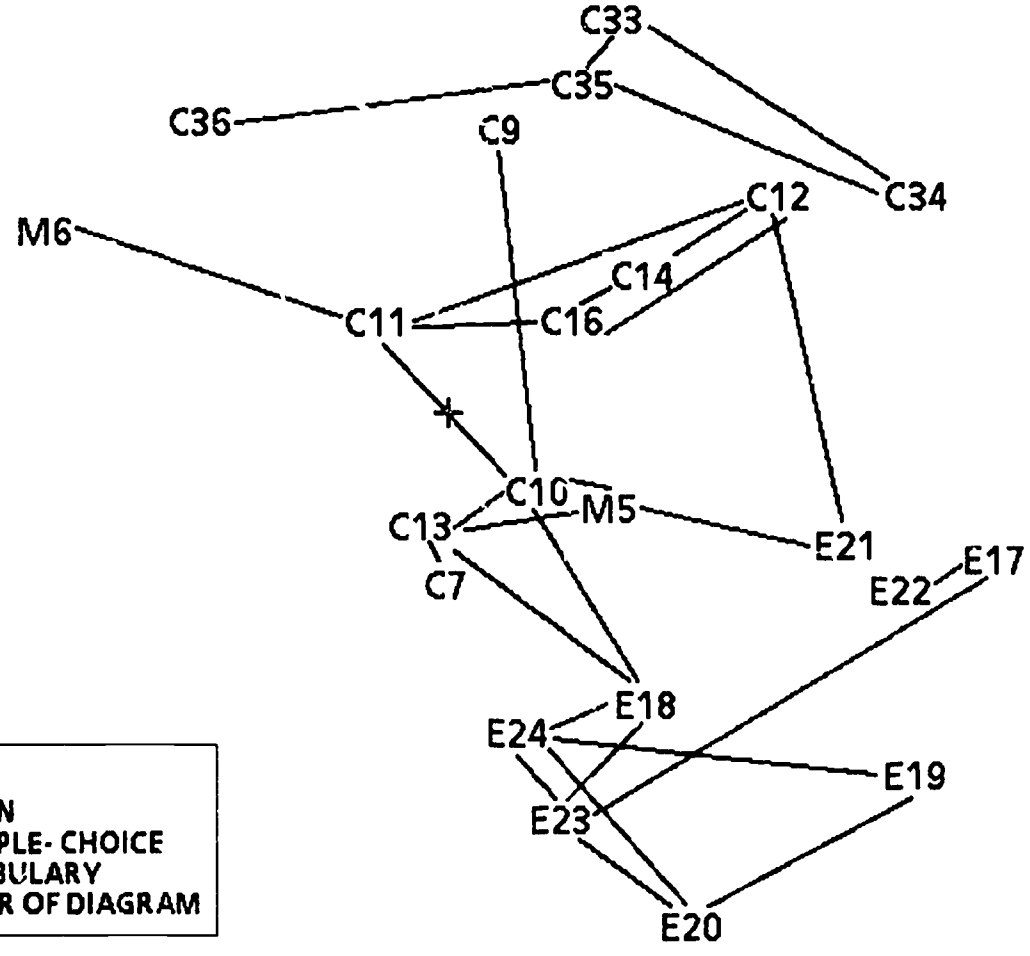


C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM

Table 4d

FIVE DIMENSIONS

NEIGHBORHOOD ANALYSIS OF 22 CORE ITEMS
2 X 1 DIMENSIONS IN 2-, 3-, 4-, AND 5- DIMENSIONS



C = CLOZE
E = ELISION
M = MULTIPLE- CHOICE VOCABULARY
+ = CENTER OF DIAGRAM