DOCUMENT RESUME

ED 390 267 FL 023 459

AUTHOR	Davidson, Fred
TITLE	Language Test Unidimensional Model Fit at Multiple Ability Levels.
PUB DATE	Feb 95
NOTE	42p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Difficulty Level; Language Aptitude; Language Proficiency; *Language Skills; *Language Tests; *Statistical Analysis; Test Construction; *Test Interpretation; Test Use

ABSTRACT

This study examined initial evidence of changes in fit to a unidimensional model for some language tests at multiple ability levels. Seven data sets were analyzed using the first phase of exploratory factor analysis: principal component eigenvalue extraction. Each data set is analyzed at varying n-sizes: whole group; random subsample; and five normally-distributed ability groups. Smoothed inter-item tetrachoric correlation coefficient matrices are used. Results suggest that restriction of range (at ability levels) yields eigenvalues that give initial evidence of poor relative fit to a unidimensional model. Whole group and random subsample matrices, however, yield better evidence of fit to a unidimensional model. Restriction of range is a possibility of operational test use, and further research is needed on precisely what kinds of factor structures underlie language tests at varying ability levels. Specifically, this study suggests that a score on a given test at one institution might not mean the same thing as the same score at another institution, if the institutions do not see the same ability range. The survey nature of the project suggests that this concern may be applicable to a wide variety of contexts of language test development and use. Contains 51 references. (Author/MSE)

Langauge Test Unidimensional Model Fit at Multiple Ability Levels

Author: Fred Davidson

Language test unidimensional model fit at multiple ability levels[1]

by

Fred Davidson
University of Illinois - English as an International Language
707 S. Mathews, Rm. 3070
Urbana, IL 61801 USA

Abstract.

This study examines the initial evidence of changes in fit to a unidimensional model for some language tests at multiple ability levels. Seven datasets are analyzed via the first phase of exploratory factor analysis: principal component eigenvalue extraction. Each dataset is analyzed at varying n-sizes: whole-group, random subsample and five normally-distributed ability groups. Smoothed inter-item tetrachoric correlation coefficient matrices are used. Results suggest that restriction of range (at ability levels) yields eigenvalues which give initial evidence of poor relative fit to a unidimensional model. Whole group and random subsample matrices, however, yield better evidence of fit to a unidimensional model. Restriction of range is a possibility of operational test use, hence further research is needed on precisely what kinds of factor structures underlie language tests at varying ability levels; specifically, this study suggests that a score on a given test at one institution might not mean the same thing as the same score at another institution, if the institutions do not see the same ability range. The survey nature of this project suggests that this concern may be applicable to a wide variety of contexts of language test development and use.

1. Introduction.

Recent years have seen a number of studies concerned with the factorial dimensionality of language test response data; more accurately, this line of research examines the degree to which a given dataset can be said to meet a unidimensional model -- a model which is an assumption of both classical and item response test theory. These studies have variously employed confirmatory factor analysis ('CFA') or exploratory factor analysis ('EFA'). Some studies have discerned multiple factors underlying language ability, and others have discerned single factors. When multiple factors were uncovered, typically there was a second-order general factor and the multiple factors were correlated. This quest for factor analytic evidence of language dimensionality seems to have come to some sort of consensus: language ability is either unidimensional, or it is multidimensional where the multiple dimensions are themselves correlated and dominated by a second-order general language ability factor. For a survey of recent factor analytic approaches to language tests, see Sawatidrakpong (1993).

Henning (1992) argued that trait modeling of language tests should distinguish the psychometric model from the psychological model. His argument is based on principal component analysis (PCA) eigenvalue extraction from simulated data; study of PCA eigenvalues is a traditional first step in determining the number of factors defining the best model for a multivariate dataset, as for example a multi-item test (see Lord, 1980: 19). Henning notes that a psychometric model, e.g. the Rasch model, is a

measurement model, whereas a psychological model is an unmeasurable model of the mind and can only be inferred from tests, and tests are always indirect. The psychometric model is of direct and immediate concern in all test construction, since unidimensionality is an assumption of both classical and latent-trait approaches to test construction. However, the psychometric model may not adequately reflect the psychological model, and what is more, detection of psychometric dimensionality can be quite difficult. Henning notes "Indeed, if one wishes to detect psychometric multidimensionality, one must carefully create the conditions under which it may be expected to occur." (Henning, 1992: 9).

If a test does not fit a psychometric unidimensional model but is used as if it does, then validity of decisions based on that test is at risk. The score user, in that case, is assuming that results of such a test indicate measurement of only one thing, and decisions about examinees might be based on that faulty (unidimensional) assumption. Richer, more complex human information -- multiple dimensions -- is lost in the score interpretation process[2]. It is this worry which has motivated the present research. This paper is concerned with situations where a score user might interpret test results as fitting a unidimensional model, when the best model for the test might be multidimensional. The present research explores a "condition under which [multiple dimensions] could be expected to occur" (Henning, *op cit*): that in which the score user does not encounter the entire ability range, as when, for example, the score user routi-

nely sees only very able students.

Previous studies of dimensionality in language testing have not examined data at multiple ability ranges, but it does generally seek to establish Henning's "conditions" favorable to multidimensionality. Bachman and Palmer (1982) designed a test battery from an avowedly multifaceted language theory based on Canale and Swain (1980), administered the battery, and used multi-trait multi-method CFA procedures to demonstrate the presence of two correlated first-order factors dominated by a second-order general language ability trait.[3] Such intentional multidimensionality was also seen in Fouly, Bachman and Cziko (1990) and Turner (1989). These studies display a confirmatory approach to detection of multidimensional language ability, and the multidimensional language ability models detected berefited from such intentionality.

Clearly, CFA differs from EFA in that the researcher imposes a theoretical model on the data and tests its closeness of fit. An alternate approach would be to set up an EFA study with available data, but intentionally alter examinee characteristics under some theoretically justified scheme intended to produce multidimensionality, and it is this approach followed in the present research. Such an approach would probably mean subdividing a test group along a categorical variable. Sawatdirakpong (1993) and Swinton and Powers (1980) did this in their studies of the effect of native language on test trait structure in the TOEFL. Both of those studies did indeed detect variability in the TOEFL trait structure dependent upon native language group.

There seems to be little point in further studies like Davidson (1988) where a large number of language test datasets are submitted to dimensionality analysis without any a priori alteration that should intentionally affect the number of traits detected; unidimensionality will probably tend to dominate the picture. To that end, this paper reports on a re-analysis of most of the Davidson (1988) datasets but under intentional conditions that may affect dimensionality, namely multiple ability groups. The research question is this: does initial PCA evidence of the item-level dimensionality vary across multiple ability groups? That is, is it possible to subdivide a large dataset -- itself tending to fit a PCA-based psychometric unidimensional model -- into smaller normally-distributed ability subgroups which appear psychometrically multidimensional?

The a priori categorical variable of ability levels is used in this study for two reasons, one minor and the other major. First, it is somewhat possible that a static language test, i.e. a single measure given at a single instant in 'interlanguage time', could be sensitive to the flux of language acquisition. Language acquisition proceeds through stages involving improvement and loss of components of language proficiency. This argument for using multiple ability groups is essentially one of applied linguistics theory.

The second argument for using ability grouping as a categorical variable for factor analysis of language data relates to operational use of test scores; this is a more realistic concern in test use, as will be argued about the author's home

institution in the Conclusion, below. Interpretation of the score of any major language test assumes interpreter has data from a test administration maximally similar to that under which the test was normed: the test must display 'norms appropriacy' (see Davidson, 1994). The distribution of observed scores by a score-using institution, e.g. placement test results at a University, may not reflect the same range of ability on which the test was constructed, and if not, then those observed scores may reflect a different composite of abilities. Test development companies wisely advocate local validity studies to determine if their product is appropriate to a given setting. This is because a test comes to have a certain 'local meaning' based on the history and memory of local ecology. It may be argued that 'a score is a score is a score' (so to speak) and is always based on a larger hypothetical population. However, the author takes exception with this view; it is epistemologically weak in the face of local use (and abuse) of major tests by educational institutions. Responsibility for proper test use must involve careful study by score users -- not only of dimensionality but of many important features of test appropriacy (see, e.g. Davidson, 1994). This paper contends that this second ecological argument is more crucial than the first linguistic argument since the state of language acquisition theory is still burdened with multiple competing theoretical perspectives (see Larsen-Freeman and Long, 1991: 227; Hatch, et al., 1990). If a dataset is psychometrically multidimensional then it may pose a threat to the interpretation of the total score, regardless of how the

dimensions might be interpreted by applied linguists.

This study presents a number of new approaches to language test dimensionality. First, it uses factor analysis of binary item-level data, itself not new to language ability modeling but certainly not widespread (see Swinton and Powers, 1980; Davidson, 1988). Second, it is a survey of a number of extant datasets, not just a single measure. Hence this study enhances generalizability of findings beyond the population and administration context of a single language test.

It must be clarified at the outset that this is a dimensionality study and not a trait structure study. In EFA terminology, the latter is only feasible when one fully extracts the factors and attempts to interpret their simple structure. The factor analyst must follow careful and justified procedures to come up with a clearly interpretable exploratory solution. That process is the focus of current research sparked by the present findings. The present study addresses the dimensionality of the tests analyzed, and it cannot make any claims about the precise nature of dimensions detected unless and until such fuller EFA is performed. This point will be addressed more fully in the concluding remarks, below.

2. Methodology.

2.1. Description of datasets.

Seven datasets from five sources are used in this study. Each is binary scored item level data, where a zero indicates an item skipped or answered incorrectly and a one indicates an item

answered correctly. Differences due to skipped and missed items were not addressed in this study because it was necessary to treat all datasets similarly, and at least two (the '87 ESLPE and Cambridge) treat skipped as incorrect when computing total scores.

Six of these datasets were analyzed in Davidson (1988). There is no evidence that the tests analyzed here were extensively revised since the 1988 project, with the exception of the UCLA ESLPE, and so a dataset from the 1992 version of that exam was added. The differences between the 1987 and 1992 UCLA ESLPE are described below.

Although clearly not every examinee reported in this study took each of the seven exams, this study does impart an exploratory survey approach to data analysis. The datasets included here represent 603 EFL language test items administered to 13,785 examinees. The data represent both ESL and EFL environments from a wide variety of test contexts, design strategies, sociopolitical concerns, examinee preparation variables, and similar considerations.

Following is a brief description of each exam:

2.1.1. The Alderson Sri Lanka Data.

The Alderson Sri Lanka data are results from the Sri Lanka National Certificate in English (NCE). This test was developed for use in Sri Lanka by Charles Alderson, Dianne Wall, and Caroline Clapham of the University of Lancaster, U.K., and colleagues. Clapham reports:

> [The NCE] is sat by people in Sri Lanka who want a Certificate in English but have not already acquired something like 'O' or 'A' levels at school. It is therefore sat by a wide range of people from school children to people in their 70s. (Caroline Clapham, personal communication)

This exam was used in a wide scale effort to engender change in the Sri Lankan English curriculum through backwash effect of the test on teaching practice (see Wall and Alderson, 1993). The data included here are Part 1 of the exam, a written mode test of reading, word completion, cloze and writing. Part 2 is not included because it is a mix of scalar and binary items, and this study examines binary scored items only.

2.1.2. The Cambridge Paper 1 Data.

The University of Cambridge Local Examinations Syndicate produces a suite of EFL tests administered worldwide for a number of purposes. The Cambridge Paper One dataset is data from a 1986 administration of the Certificate of Proficiency in English (CPE), in particular, the 40 items from subtest/paper one. Paper One is an exam of 25 English usage/vocabulary items followed by fifteen comprehension items on three reading passages (five items per passage).

2.1.3. The Fall 1987 UCLA ESLPE

The Fall 1987 UCLA English as a Second Language Placement Exam (ESLPE) is a 150-item multiple choice test of English, covering grammar, listening, reading, error detection, and vocabulary. It was developed following norm-referenced procedures using item banking. This author worked on that exam.

2.1.4. The Fall 1992 UCLA ESLPE

The Fall 1992 UCLA ESLPE is a criterion-referenced 70-item test which replaced the 1987 NRM ESLPE in 1990. The 1992 ESLPE is built along criterion-referenced language test development guidelines very similar to those sketched by Lynch and Davidson (1994) and influenced by the work of W.J. Popham (1978, 1990: Chapter 9). This criterion-referenced development process behind the 1992 UCLA ESLPE involves consensus among test administrators and teachers in the UCLA ESL course sequence. The consensus is achieved through iterative trial-and-error of test specifications. Brian Lynch (personal communication) of the UCLA faculty reports that by 1992, the ESLPE is designed to be intentionally quite different from the 1987 exam, and further, that it is designed to represent multiple facets of language ability. The Fall 1992 exam tests reading and listening.

2.1.5. Jones Southam Literacy Tests, Versions 1 and 2.

Two datasets are provided by Stan Jones of Carleton University, Ottawa, Canada: the Southam Literacy Tests, Version 1 and Version 2. These are tests intended to ascertain the functional English literacy of the examinee. The examinees are a mixture of native speakers of French and English, though both tests are of English. The two tests were developed together. Each contains reading items such as repair notices, driver's licenses, and other real-world literacy skills. (Stan Jones, personal communication.)

2.1.6. 1985 TOEFL, random subsample.

This project also includes a sample TOEFL dataset. The TOEFL is given worldwide as a screening measure, most typically for entry to U.S. universities. At the time these data were collected, approximately 500,000 persons per year took the official TOEFL (Grant Henning, personal communication). This dataset is a random sample of 5000 exam takers from the 1985 TOEFL group.

2.2. Analytical Procedures.

For each dataset, the following steps were followed:

2.2.1. Creation of subgroups.

Five near-normally distributed ability subgroups were constructed using an algorithm written in SAS (SAS Institute Inc., 1988a)[4]. Normality, an assumption of factor analysis, is therefore not a variable under analysis in this study; that is, the effect of non-normality on item-level factor analysis is not examined.

The procedure to extract the normal subgroups was as follows. First, a procedure was run to determine the whole group total score 10th, 30th, 50th, 70th and 90th percentiles. Next, the raw score corresponding to each of those percentiles was used in a computer routine as a mean value for each subgroup. Following the suggestion of De Jong (1990 and personal communication), the standard deviation for each subgroup was set at one-half the standard deviation of the whole group, in order to avoid creation of extremely flat subgroup distributions. The computer algorithm relied on the SAS random/normal number function 'rannor' (SAS Institute Inc., 1988a: 90), hence the resulting group usually had

slightly different observed mean and standard deviation. The n-size for each ability subgroup was fixed at one-fifth that of the whole group; once the algorithm reached that n-size, the subsetting program terminated. By convention hereafter, the '1st' group will be the lowest-ability group and the '5th' group the highest ability.

Five subgroups were extracted rather than six as done by De Jong (1990). This was done to ensure that group 3 spanned the observed mean of the whole group, since if the whole group is normal then the 50th percentile would represent the raw score mean. This would allow analysis of a group with a restricted range but a similar mean to the whole group, a possibility in operational test use. Furthermore, for maximum comparability across datasets, it would not have been feasible to extract much over five subgroups because the total n-sizes were sometimes small, and some of the subgroups would have been fairly small.

As a check, another subgroup was extracted for each dataset by randomly selecting one-fifth of the examinees from the whole group. This group, called the 'random subsample', was intended to mirror the whole group to see if the reduction in the n-size alone affected results. This random subsample was intended to present effectively the same mean, range and standard deviation as the whole group but at an n-size identical to that of the five ability subgroups.

Hence, for each of seven datasets, there are seven datasets under analysis:

-the whole group

-random subsample (an n-size reduction check, same range, mean and s.d. as whole group)
-1st subgroup (lowest ability)
-2nd subgroup
-3rd subgroup (with a mean very similar to the whole group, but a smaller range and standard deviation)
-4th subgroup
-5th subgroup (highest ability)

This yielded 49 datasets in the entire study.

2.2.2. Descriptive statistics.

For each of the 49 datasets, complete descriptive statistics were run, all in SAS, with the exception of the computation of the alpha reliability coefficient, done in SPSS-PC+ (Norusis/SPSS, Inc., 1988). The descriptives included skewness and kurtosis, because a key goal of this project is near-normal ability subgroups. Without near-normality of such subgroups, factor analysis to determine potential dimensionality may be overly influenced by distributional shape.

2.2.3. Principal Components Analysis.

For each of the 49 datasets, inter-item tetrachoric coefficients were computed by two methods and a principal components analysis (PCA) run to check dimensionality by examining the eigenvalues. Unsmoothed coefficients were computed using the PRELIS program, a preprocessor to the LISREL software (Jöreskog and Sörbom, 1988). Smoothed tetrachorics were computed using TESTFACT, an item-level factor analytic software package (Wilson et al., 1991).

'Smoothing' refers to the use of an interpolative mathematical procedure to reduce the effect of extreme values of inter-item correlations, as when items have extremely high or extremely

low p-values. Without smoothing, tetrachoric matrices tend to be singular, and hence not amenable to the later stages of some forms of factor analysis. Hence, in using both unsmoothed and smoothed tetrachoric matrices, this study provides a comparison of more technically advanced tetrachoric computation in the form of smoothed coefficients (TESTFACT) to a more widely available rough-and-ready coefficient matrix (PRELIS, which is now a sub-routine of the common mainframe program, SPSS). For further discussion of TESTFACT and the issue of smoothing, see Muraki, 1984; Bock et al., 1988; Lawrence and Dorans, 1987; Wothke, 1993). For discussion of the computation of the tetrachoric coefficient see Harris, 1988; Divgi, 1979. In the interest of presenting results most similar to current thinking in item-level factoring, only the smoothed analyses are given in tables here, but some commentary on comparability with unsmoothed coefficients will be provided below.

PRELIS was unable to compute tetrachorics for any inter-item pair where either item had zero variance. Several such items were detected in this study, in all cases items passed by all students. To ensure maximum comparability between both smoothed and usmoothed matrices, those items were eliminated in all 98 PCA runs. Technically, however, TESTFACT is able to compute tetrachorics for zero-variance items.

TESTFACT can also provide item-level factor analysis which is corrected for guessing, using c-parameter estimates from an IRT program like BILOG (Mislevy and Bock, 1990). However, in order to correct for guessing, it is necessary to commit a TES-

TFACT run to a particular number of factors to extract, and in this study no such commitment was assumed, since this study sought to find evidence of a potential number of factors underlying the observed tetrachoric matrices. For further discussion of guessing correction in the factor analysis of tetrachoric matrices, see Bock et al., 1988; Lawrence and Dorans, 1987.

The tetrachoric coefficient has been suggested as an alternative to the inter-item variant of the Pearson coefficient, the Phi value, to avoid extraction of a difficulty factor in factor analysis (Bock et al., 1988: 261). However factor analysis of tetrachoric matrices is somewhat controversial. Lord (1980: 19) advocates use of tetrachorics but Sawadirakpong (1993) contends they may be problematic. No other binary inter-item coefficient has emerged resoundingly in the literature to replace the tetrachoric in factor analysis of binary data.

Each tetrachoric matrix was saved as an external file and passed back into SAS, which was selected because it has the data-handling capabilities necessary for the factor analyses and production of output eigenvalue datasets. In SAS, an unweighted least squares (ULS) principal components analysis (PCA) was run to obtain initial eigenvalues. The input unsmoothed (PRELIS) tetrachoric matrix had unities on the diagonal in the PCA, whereas the input smoothed (TESTFACT) tetrachoric matrix had squared multiple correlations (SMCs) on the diagonal. Hence, the smoothed PCA represents a more technically accurate approach to factor analysis since error is not a component of the input matrix. SAS (SAS Institute, Inc., 1988b: 453) refers to PCA with

SMCs on the diagonal as 'principal factor analysis'; here it is simply referred to as PCA with SMCs.

The eigenvalues from each matrix were saved to an external file and passed back into another SAS program to compute the percentage of each eigenvalue out of the total of all eigenvalues.[5] The first twenty eigenvalues were used to produce scree plots. A scree plot is a graphic representation of eigenvalues to aid the researcher in deciding on the most appropriate number of factors to extract. The vertical axis of a scree plot is the magnitude of the eigenvalue, and the horizontal axis is its sequential cardinal number. Cattell proposed this judgmental tool to aid in the classical problem of the number of factors in a matrix of correlations. (For background, see Cattell, 1966; Cattell and Jaspers, 1967; Cattell and Vogelman, 1977; Cattell, 1978: Chapter 4; Horn and Engstrom, 1979; Hakistan et al., 1982; Zwick and Velicer, 1982. Additionally, Berger and Knol, 1990, provide a good overview of the number-of-factors problem in EFA). The name 'scree' derives from geography and refers to the scree left by a retreating glacier. That rubble forms a distinctive descending pattern which ultimately becomes quite level. Beyond the point where the plot becomes level, often called an 'elbow', very little or no information is to be gained by successive numbers of factors. Scree plot interpretation is admittedly something of an art. Ideally, it involves extraction of a number of factors corresponding to the eigenvalue numbers nearest the clearest elbow of the plot, although Cattell describes a 'test' involving tangent lines. Scree plots are best as a rough indica-

tor of a range of number of factors to extract. The factor analyst then performs full factor extractions and determines which final factor pattern meets interpretability criteria such as simple structure. Davidson (1988), in analyzing the whole group data represented here, claimed that the overwhelming number of L-shaped scree plots were indicators of general unidimensionality of the datasets surveyed. Regrettably, space limitations prevent printing all scree plots here, but some representative samples are provided. All scree plots in this study were created using Quattro Pro (Borland, Inc., 1989).

Finally, eigenvalue drop tables were constructed via another SAS program and editing. Eigenvalue 'drop' is defined as the result of dividing the first eigenvalue by the second eigenvalue. It provides a way to summarize across the 49 scree plots. If the DROP value is extremely large, then that is initial evidence that the dataset may be unidimensional; a large drop value would translate into a tendency for a L-shaped scree plot, indicative of possible unidimensionality.[6] To allow comparison across the seven data sources, the drop analyses employ eigenvalue percentages rather than raw values. It should be noted that eigenvalue drop was one of the dimensionality indices examined by Hattie (1984, 1985), as is the alpha reliability coefficient calculated in this study. For recent related literature on the issue of unidimensionality indices, see Knol and Berger, 1991; DeAyala and Hertzog, 1991; Roznowski et al., 1991.

The present study is a cross-sectional analysis of 49 datasets at the earliest factor analysis step: PCA. No further

factoring was performed.

3. Results.

3.1. Descriptive Statistics.

Insert Table 1 about here.

Table 1 presents descriptive results for all 49 datasets. The normal-subgroup extraction algorithm appears to be working fairly well. The skewness and kurtosis of the ability subgroups tend not to be extreme. Furthermore, the means and ranges indicate that the subgroups are covering the entire range of the whole group in a stairstep fashion; each subgroup presents a higher mean, similar standard deviation, and relatively higher segment of the range of the whole group. Admittedly, it is quite difficult to achieve a perfectly normal ability subgroup with clear monotonic relation to the range of the whole test; note, for example, that the 3rd, 4th and 5th subgroups of the 1992 UCLA top out at or near the range of the whole group. This is due to the fact that the subgroup extraction algorithm is at the mercy of the n-size and distributional characteristics of the whole group. Finally, the standard deviation of the subgroups does not generally stay at the desired value, 1/2 that of the whole group. This is also probably due to distributional characteristics of the whole group as well as the precise nature of the SAS 'rannor' normal function (SAS Institute Inc., 1988a: 90).

Throughout the datasets, the reliability coefficients of the ability subgroups are generally lower than that of the whole

group. Further, the random subsamples tend to have very nearly the same mean, standard deviation, range, and reliability coefficient as the whole group, whereas the ability subgroups do not. Clearly, the range-reduction of the ability subgroups has affected the reliability coefficient, as is to be expected. Both correlation and reliability coefficients are expected to drop when range is restricted, a point to which this paper will return.

3.2. PCA

Insert Table 2 about here.

Table 2 presents the first of two styles of eigenvalue drop analysis. The percentage first, second, and drop values are presented for each dataset, in a descending sort. It is interesting to note that for every dataset the highest drop value is either for the whole group or the random subsample. Furthermore, there is not much consistency across datasets as to the ordering of the drop values beyond that for the whole group and random subsamble. For example, Table 2 shows that the 5th Alderson subgroup ability subgroup has the largest drop value beyond the whole group and random subsample, whereas for the Cambridge Paper 1 exam the same position is occupied by the 1st ability group. There may be a tendency for either the 5th or 1st to have the highest drop value for any ability subgroup, particularly for the smoothed analysis. Generally, however, it is striking that Clearly, the PCA results for the ability subgroups differ from

the whole group or the random subsample.

Insert Table 3 about here.

Insert Figures 1 through 4 about here.

In Tables 3, all 49 datasets are presented ordered on a descending sort based on eigenvalue drop. This presentation is different than that in Table 2 -- here, comparison is possible across datasets. It must be emphasized that Table 3 does <u>not</u> claim, for example, that the 1985 TOEFL whole group is unidimensional whereas the 1992 UCLA ESLPE 3rd ability group is not (ranks 11 versus 39 in Table 3). Rather, the PCA results are intended to address this question: in which situations would an exploratory factor analyst be more motivated to seek a multidimensional factor extraction? For datasets at the top of Table 7 there would not be much motivation to do so. That contention is further supported by examination of some representative scree plots from the top of Table 3 versus some from the bottom. For example, Figure 1 (rank 1 in Table 3) is a classically unidimensional scree plot. The analyst might extract a 2-factor and maybe a 3-factor solution to check interpretability and simple structure, but would not be much motivated to go further, and s/he would not be surprised if neither the 2- or 3-factor solution appeared substantively meaningless, leaving a 1-factor solution as most tenable. On the other hand, Figure 2 (rank 46 in Table 3) would motivate extraction of 4, 5 and 6 factors, if not more. Indeed, the scree 'elbow' in Figure 2 is not easy to

detect. Generally, Figure 2 does not motivate a one-factor extraction. Figure 3 (rank 47 in Table 3) is even harder to interpret. And finally, to clarify the role of range in these analyses, it should be noted that unidimensional-appearing scree plots are not always associated with the whole group or random subsample. Figure 4 (rank 13 in Table 3) would not motivate a large number of factor extractions, though there is some indication of a possibly graphically spurious drop at about the 14th eigenvalue.

Finally, there was little difference between smoothed and unsmoothed results. Unsmoothed tetrachorics produced tables analogous to Table 2 and 3 with datasets ranked in very nearly the same order. In fact, the Spearman rank order correlation between Table 3 and its unsmoothed counterpart was +.91. Since smoothed coefficients produce matrices which are more amenable to later stages of EFA, smoothing seems justified.[7]

4. Discussion.

In summary, the results of this project suggest that normal or near-normal language test ability subgroups display different PCA-based dimensionality evidence than the whole group. They also display lower alpha reliability coefficients. Randomly extracted subgroups of the same range as the whole group generally resemble the whole group.

Mathematically, these results are not surprising. Range restriction via simple sample selection cutoff can also reduce correlation coefficients, as noted by Shavelson (1981: 210-211;

it can also cause reliability to drop, a phenomenon noticed here as well). Although this project did not simply truncate distributions to obtain subgroups (great care was taken to ensure the normal distribution of the ability subranges), a single dominant factor should also be less evident in a restricted range than in a full range (see Comrie, 1973: 201-202; Gorsuch, 1983: 342-346). This suppression of a dominant single factor is because the input variables must have uniform and high intercorrelations among the input variables to meet a unidimensional factor model. This paper has found precisely what should happen, and that fact is acknowledged.

There are two reasons why research at restricted ability ranges must proceed, despite such mathematical predictability of the results. The first argument is that restricted ranges may exist in practice. Regardless of the predictability of altered factor structures at ability ranges, if a score-using institution is encountering such restricted ranges, then the factor structure might change. More generally, the issue of predictable mathematical phenomena does not vitiate the existence of those phenomena. Score users should accommodate whatever potential threat such phenomena pose to test interpretation, if the phenomena exist.

A second argument against the predictability of range-restriction, and in favor of future research along this line, is that some of the ability subgroup ranges observed in this study are not extremely reduced. The entire normal-subgroup extraction process in this paper is a prisoner to the distribution of each whole group and the design characteristics

of each test. Some tests in this study have relatively restricted ranges to begin with, e.g. the Cambridge or either Jones' Southam Literacy test. When working with a total test range of 40 (Cambridge) or 38 or 39 (Jones), splitting the distribution into five ability subgroups tended to produce subgroups with relatively wide ranges relative to the whole group (Table 1). Yet the ability subgroup results for those three datasets are distributed pretty evenly in Table 3. Further research on extraction of normally distributed ability subgroups is clearly merited.

It could also be claimed that this research has served to confirm the mathematical prediction of altered factor structures at restricted ability ranges, and in that sense, this paper has relevance to literature on factor analysis.

Perhaps there <u>are</u> clear factor analytic differences at multiple ability levels in the tests analyzed here. To detect such differences, it would be necessary to run complete EFAs at each potential number of factors, rotate to a terminal solution, and check the interpretability and simple structure.[8] It should be noted that a SAS unweighted least squares PCA extraction was used here, which does not abort with a singular matrix. The more powerful maximum likelihood (ML) method would have aborted in such a situation. Therefore, if further full EFAs were run and the ML extraction selected, only the smoothed coefficients could be used. This is somewhat problematic in that this study corroborates that smoothing can affect slightly the pattern of eigenvalues (see the discussion of this point in Sawatirakpong, 1993).

Furthermore, fixing skipped items as wrong (as this study did) can heighten unidimensionality (Lawrence and Dorans, 1987). In short, at least two difficult decisions must be made before further EFA extractions are pursued with these datasets: handling of skipped items and the method of extraction. The latter concern is epistemologically difficult since committing to smoothed coefficients effectively implies alteration of input data. Ideally, further EFA extractions should examine the effect of each of these problems. While such research is feasible, it would have been quite time-consuming and was beyond the scope of the present study.

5. Conclusion

This study has demonstrated initial evidence for concern of the stability of item-level dimensionality across multiple ability levels in a survey of seven language tests from five sources. Earlier work (Davidson, 1988) on the whole group datasets replicated here may have indicated clear unidimensionality, but that interpretation must be tempered by the requirement of analyzing the entire examinee range.[9] This topic reflects a potential threat to test validity in the modern sense of validity for an intended particular use (see APA/AERA/NCME, 1985; Messick, 1989; Shepard, 1993). Norm-referenced testing relies on stability of rank-based judgments. The dimensionality of a test may vary as a function of the particular examinee group -- here ability-based grouping, and such judgments could be faulty, which is a threat to test validity.

In practial terms, the findings of this paper could suggest a scenario such as the following. Alpha and Beta Universities are in an English-speaking country. They both use a major ESL test as an entry screening device: the 'WTE' (World Test of English). At Alpha U, only the upper end WTE students routinely apply; Alpha tends to see a normally-distributed cohort which is something like the '5th' group examined in the tests in this paper. On the other hand, Beta U tends to receive intermediate applicants, something like a 'group 3' cohort on the WTE. The results of this project suggest that a given score on the WTE might not mean the same thing at Alpha U as it does at Beta U. In both cases, these findings suggest, the WTE score could be interpreted as having better relative fit to a multidimensional model, but if the WTE behaves as did some tests examined here (if the eigenvalues were different for WTE groups 3 and 5), then <u>different dimensions</u> might best define the composite trait structure at each university.

Perhaps there is a third university -- Gamma -- which also uses the WTE. Perhaps it regularly sees a normally distributed WTE applicant pool covering the entire ability range. Of the three universities, Gamma U can (relatively) best assume that its WTE results fit a unidimensional model. Gamma U would be in the best position to regularly interpret the WTE as measuring a single trait.

In all three universities, the number of traits assessed by the WTE is crucial. Each university must use the WTE information to make entry decisions, and those decisions would be enhanced by

knowing precisely what the WTE measures. Further, WTE evidence might figure into ESL program design or the mix of ESL and non-ESL course load permitted. To the extent that the WTE does figure strongly in the three infrastructures, perhaps somewhat different ESL classes would evolve at each university.

These scenarios reflect the situation at this author's home institution. It does not encounter the entire TOEFL ability range in its ESL service courses. The low end of TOEFL-scorers either (1) do not apply to the author's university because they know of our campus cutoff, or (2) apply and are rejected by that cutoff. Furthermore, unless students voluntarily enroll in ESL (not common) the ESL service course stream never encounters students above a maximum value at which the local ESL test is waived.[10] In short, the international students who arrive at the author's campus and do indeed take ESL service probably represent something like the 4th TOEFL ability group above. Historically, the institution's ESL test has come to serve precisely the correct function; it gives information that is validated against the English needs of the campus using criterion-referenced test construction technology. The results of this study have shown that the author's institution has intuitively followed the correct course: this paper further supports the classic warning that a test should be used for the purpose for which it was intended. It provides further evidence that the TOEFL should not be used for placement decisions. However, this is new evidence for two reasons: (1) it shows a new reason not to use TOEFL as a placement tool: its dimensionality may vary as a function of ability

group, and (2) more importantly, it shows that TOEFL may not be alone in this phenomenon: other major world English tests display tne same tendency.

In none of these four universities -- the hypothetical Alpha, Beta, or Gamma, or the real author's institution -- does there yet exist a fully extracted EFA of item-level results for each ability level. That is the next logical step. The present project has served to indicate that such a step should be taken, based on initial PCA eigenvalue evidence. As noted above, it is imperative that fully-interpreted factor analysis now be carried out at multiple ability levels, at the whole-group, and at the random subset. It is also imperative that the survey flavor of this project be continued. Surveying many tests in factor analysis allows generalizability across multiple testing contexts, and such generalizability should be of interest as EFA analyses begin to identify actual trait structures at multiple ability levels.

ACKNOWLEDGEMENTS

The author gratefully thanks the supplier of each dataset.

NOTES

[1] This research was supported by the University of Illinois Research Board, grant numbers 1-2-69941 and 1-2-68677.

[2] Davidson (1994) reasons that one type of information lost in the construction of highly unidimensional language tests is

richness of sociolinguistic variety. As tests are developed for high reliability and a single common factor, items which tap less frequent forms are selected against. In the words of Henning (*op cit*) one has not "carefully create[d] the conditions under which [multidimensionality] might occur."

[3] In his extensive review of cognitive ability factor analysis, J.B. Carroll (1993: 191) notes that his EFA re-analysis of the Bachman and Palmer (1982) data did not reveal precisely the same factor structure claimed by Bachman and Palmer, although separable factors were uncovered by Carroll. That does not vitiate the intentional multidimensionality of the Bachman and Palmer study. The key issue is that Bachman and Palmer set out to find multiple factors and found them. Carroll found them also. Any debate rests in how those factors are interpreted.

[4] This SAS algorithm is available on request from the author.

[5] The first twenty eigenvalues for all 49 datasets is available on request from the author.

[6] It must be emphasized that a large eigenvalue drop does not always translate into a perfectly L-shaped scree plot, a shape which would suggest unidimensionality. The scree elbow can have a later bend or curve, which would motivate further EFA at several numbers of factors. Eigenvalue drop is a rough-and-ready statistic.

[7] Precise data on comparison of the smoothed and unsmoothed results is available on request from the author.

[8] Previous studies of the TOEFL, e.g. Davidson (1988), have uncovered a two-factor structure comprised of the listening test and the balance of the exam. It is possible that such a factor structure is the best fit for some or all of the seven TOEFL analyses here. But as noted, to determine this precisely, a fully interpreted EFA would need to be run.

[9] Citation of the Davidson (1988) study as evidence of dimensionality should note that it was a project which examined the entire ability range of each dataset reported there.

[10] The overall campus admission minimum is 550. Students admitted between 550 and 607 must also take the local ESL test. Some departments have higher upper maxima (e.g. Linguistics or the author's home unit).

REFERENCES.

American Psychological Association (APA), American Educational Research Association (AERA, and the National Council on Measurement in Education (NCME). 1985. Standards for Educational and Psychological Testing. Washington, DC: APA.

Bachman, L.F. and A.S. Palmer. 1982. The construct validation of some components of communicative proficiency. TESOL Quarterly 16:4, 449-466.

Berger, M.P.F. and D.L. Knol. 1990. On the assessment of dimensionality in multidimensional item response theory models. Research Report 90-8. Twente University, Enschede (Netherlands), Dept. of Education. Eric Document Reproduction Service Number ED 329 584.

Bock, R.D., R. Gibbons and E. Muraki. 1988. Full-information factor analysis. Applied Psychological Measurement 12:3, 261-280.

Borland International, Inc. 1989. Quattro Pro User's Guide. Scotts Valley, CA: Borland International, Inc. Computer software.

Canale, M. and M. Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1:1, 1-47.

Cattell, R.B. 1966. The scree test for the number of factors. Multivariate Behavioral Research 1:2, 245-276.

Cattell, R.B. 1973. The Scientific Use of Factor Analysis in Behavioral and Life Sciences New York: Plenum Press.

Cattell, R.B. and J.J. Jaspers. 1967. General plasmode no. 30-10-5-2 for factor analytic exercises and research. Multivariate Behavioral Research Monographs #3.

Cattell, R.B. and S. Vogelman. 1977. A comprehensive trial of the scree and KG criteria for determining the number of factors. The Journal of Multivariate Behavioral Research 12:3, 289-325.

Carroll, J.B. 1993. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge, U.K.: Cambridge University Press.

Comrie, A.L. 1973. A First Course in Factor Analysis. New York: Academic Press.

Davidson, F.G. 1988. An exploratory modeling survey of the trait structures of some existing language test datasets. Unpublished Ph.D. dissertation, University of California at

Los Angeles. Dissertation Abstracts International number DA8815771.

----------. 1994. Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms. Language Testing (forthcoming)

----------. 1994. "The Interlanguage metaphor and language assessment" World Englishes 13:3, pp. 377-386.

De Ayala, R.J. and M.A. Hertzog. 1991. The assessment of dimensionality for use in item response theory. Multivariate Behavioral Research 26:4, 765-792.

De Jong, J.H.A.L. 1990. Test dimensionality in relation to student proficiency. Paper presented at the 12th Annual Language Testing Research Colloquium, San Francisco, March, 1990.

Divgi, D.R. 1979. Calculation of the tetrachoric coefficient. Psychometrika 44:2, 169-172.

Fouly, K.A., L.F. Bachman and G.A. Cziko. 1990. The divisibility of language competence: a confirmatory approach. Language Learning 40:1, 1-21.

Gorsuch, R.L. 1983. Factor Analysis, Second Edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hakistan, A.R., W.T. Rogers and R.B. Cattell. 1982. The behavior of number-of-factors rules with simulated data. Multivariate Behavioral Research 17:2, 193-219.

Harris, B. 1988. Tetrachoric correlation coefficient. Encyclopedia of Statistical Sciences, Volume 9 New York: Wiley. 223-225.

Hatch, E., Y. Shirai and C. Fantuzzi. 1990. The need for an integrated theory: connecting modules. TESOL Quarterly 24:4, 697-715.

Hattie, J. 1984. An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research 19:1, 49-78.

----------. 1985. Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement 9:2, 139-164.

Henning, G. 1992. Dimensionality and construct validity of language tests. Language Testing 9:1, 1-11.

Horn, J.L. and R. Engstrom. 1979. Cattell's scree test in relation to Bartlett's chi-square test and other observa-

tions on the number of factors problem. Multivariate Behavioral Research 14:3, 283-300.

Jöreskog, K.G. and D. Sörbom. 1988. PRELIS: A Program for Multivariate Data Screening and Data Summarization. A Preprocessor for LISREL. Chicago, IL: Scientific Software International. Computer software.

Knol, D.L. and M.P.F. Berger. 1991. Empirical comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research 26:3, 457-477.

Lawrence, I.M. and N.J. Dorans. 1987. An assessment of the dimensionality of the SAT-Mathematical. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC, April 21-23, 1987. Eric Document Reproduction Service Number ED 282 898.

Larsen-Freeman, D. and M. Long. 1991. An Introduction to Second Language Acquisition Research. London: Longman.

Lord. F.L. 1980. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lynch, Brian and Fred Davidson "Criterion-referenced language test development: linking curricula, teachers and tests." TESOL Quarterly 28:4, pp. 727-743. (refereed)

Messick, S. 1989. 'Validity'. in R.L. Linn (Ed.) _Educational Measurement, 3rd. Edition_. New York: NCME/ACE - MacMillan, pp. 13-103.

Mislevy, R.J. and R.D. Bock. 1990. BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models. Chicago, IL: Scientific Software International. Computer software.

Muraki, E.J. 1984. Implementing full information factor analysis: TESTFACT program. Paper presented at the Annual Meeting of the Psychometric Society, San Antonia, Texas, November 1-3, 1984. Eric Document Reproduction Service Number ED 260 094.

Norusis, M.J. / SPSS, Inc. 1988. SPSS PC+ Advanced Statistics V2.0 for IBM PC/XT/AT and PS2. Chicago: SPSS, Inc. Computer software.

Popham, W.J. 1978. Criterion-Referenced Measurement. Englewood Cliffs, New Jersey: Prentice-Hall.

----------. 1990. Modern Educational Measurement: A Pratitioner's Perspective. Englewood Cliffs, New Jersey: Prentice-Hall.

Roznowski, M., L.R. Tucker and L.G. Humphreys. 1991. Three approaches to determining the dimensionality of binary items. Applied Psychological Measurement 15:2, 109-127.

SAS Institute Inc. 1988a. SAS Language Guide for Personal Computers, Release 6.03 Edition. Cary, North Carolina: SAS Institute Inc. Computer Software.

----------. 1988b. SAS/STAT User's Guide, Release 6.03 Edition. Cary, North Carolina: SAS Institute Inc. Computer software.

Sawatdirakpong, S. 1993. Native Language and Differential Dimensionality of English as a Second Language Test Proficiency: An Exploratory Study Unpublished Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Shepard, L.A. 1993. Evaluating test validity. in L. Darling-Hammond (Ed.) Review of Research in Education 19. Washington, DC: AERA, 405-450.

Swinton, S.S. and D.E. Powers. 1980. Factor Analysis of the Test of English as a Foreign Language for Several Language Groups. TOEFL Research Reports Number 6. Princeton, New Jersey: Educational Testing Service.

Turner, C. E. 1989. The underlying factor structure of L2 cloze test performance in Francophone, university-level students: causal modeling as an approach to construct validation. Language Testing 6:2, pp. 172-197.

Wall, D. and J.C. Alderson. 1993. Examining washback: the Sri Lankan Impact Study. Language Testing 10:1, 41-70.

Wilson, D.T., R. Wood, and R. Gibbons. 1991. TESTFACT: Test Scoring, Item Statistics and Item Factor Analysis. 386/486. Chicago, IL: Scientific Software International. Computer software.

Wothke, W. 1993. Nonpositive Definite Matrices in Structural Modeling. in Bollen, K.A. and J.S. Long. (Eds.) Testing Structural Equation Models. Newbury Park, CA: Sage Publications.

Zwick, W.R. and W.F. Velicer. 1982. Factors influencing four rules for determining the number of components to retain. Multivariate Behavioral Research 17:2, 253-269.

Table 1. Descriptive statistics: all datasets.

Test	Dataset	n	mean	s.d.	range	alpha	skew.	kurt.
	Whole	1014	79.14	14.25	52-117	.904	0.428	-0.665
	ran sub	203	79.31	14.81	54-113	.912	0.465	-0.722
Alderson	1st,low	203	70.25	7.06	54-85	.571	0.238	-0.487
Sri Lanka	2nd	203	74.12	9.32	54-92	.763	0.148	-1.000
Part 1	3rd	203	78.65	11.43	58-101	.850	0.171	-1.008
(k=120)	4th	203	85.10	11.18	67-110	.854	0.170	-1.039
	5th, hi	203	94.47	9.57	80-117	.832	0.336	-0.981
	Whole	3886	22.38	7.97	0-40	.886	-1.121	1.521
	ran sub	777	22.52	7.78	0-39	.880	-1.061	1.561
Cambridge	1st,low	777	21.08	6.68	0-30	.827	-1.441	2.501
Paper 1	2nd	777	23.32	5.52	7-36	.744	-0.124	-0.267
(k=40)	3rd	777	23.67	5.65	10-40	.758	0.154	-0.313
	4th	777	23.92	5.42	13-40	.737	0.257	-0.381
	5th, hi	777	24.65	4.86	17-40	.675	0.469	-0.294
	Whole	844	101.46	23.99	2-140	.959	-1.152	1.632
	ran sub	169	100.17	25.20	2-140	.963	-1.271	2.081
UCLA	1st,low	169	80.81	17.36	35-107	.903	-0.525	-0.347
1987	2nd	169	103.14	15.06	62-134	.893	-0.631	0.280
ESLPE	3rd	169	110.24	14.19	78-140	.892	-0.111	-0.622
(k=150)	4th	169	112.18	13.80	83-140	.888	-0.026	-0.755
	5th, hi	169	114.96	12.00	94-140	.860	0.146	-0.774
	Whole	670	51.07	10.81	0-69	.905	-0.963	1.270
	ran sub	134	49.78	11.21	3-67	.910	-1.108	1.812
UCLA	1st,low	134	45.00	6.49	21-54	.675	-0.881	0.730
1992	2nd	134	53.27	7.32	33-65	.800	-0.412	-0.430
ESLPE	3rd	134	55.11	7.24	39-68	.812	-0.306	-0.657
(k=70)	4th	134	55.76	6.71	43-68	.786	-0.138	-0.884
	5th, hi	134	58.54	5.24	49-69	.697	0.077	-0.964
	Whole	1185	25.61	8.48	0-38	.929	-1.268	1.205
	ran sub	237	26.01	7.90	0-37	.918	-1.375	1.845
Jones	1st,low	237	20.34	7.48	2-28	.896	-1.155	0.390
So. Litr.	2nd	237	27.21	6.10	11-37	.861	-0.628	-0.305
Ver. 1	3rd	237	28.59	5.30	16-38	.825	-0.456	-0.833
(k=38)	4th	237	29.03	4.93	19-38	.802	-0.400	-0.939
	5th, hi	237	30.01	4.04	22-38	.718	-0.333	-0.929
	Whole	1186	26.66	9.11	0-39	.935	-1.124	0.609
	ran sub	237	26.91	8.67	1-39	.927	-1.100	0.588
Jones	1st,low	237	18.69	7.99	0-28	.900	-0.760	-0.442
So. Litr.	2nd	237	28.83	6.27	13-39	.859	-0.730	-0.162
Ver. 2	3rd	237	29.64	5.71	16-39	.836	-0.633	-0.405
(k=39)	4th	237	30.72	4.73	20-39	.772	-0.464	-0.599
	5th, hi	237	31.47	4.03	23-39	.700	-0.284	-0.808

Table 1. Descriptive statistics: all datasets.
[Continued]

	Whole	5000	93.07	28.85	7-146	.970	-0.363	-0.458
	ran sub	1000	92.75	29.23	26-146	.971	-0.361	-0.519
1985	1st, low	1000	70.50	17.42	23-101	.900	-0.471	-0.235
TOEFL	2nd	1000	87.21	18.96	40-127	.923	-0.145	-0.694
(k=146)	3rd	1000	95.29	19.60	56-143	.933	-0.014	-0.762
	4th	1000	104.03	18.29	70-145	.930	0.183	-0.787
	5th, hi	1000	115.39	14.82	91-146	.912	0.212	-1.024

Notes on Table 1:

-All values reflect the original raw score metric. No converted or transformed scores are presented.

-For Cambridge Paper 1, there were 234 persons with an observed total score of zero, which affected the skewness and kurtosis of the whole group results and the first (low) subgroup.

-The number of items given for each test (k) is as the test is published and as supplied in the dataset. Each dataset was full and complete (all persons measured on all items). Later analysis required that some items be dropped due to zero variance, because in PRELIS tetrachoric coefficients could not be calculated for any item-item pair where one item had zero variance. In all cases, zero-variance items were passed by all persons. Table 2 gives a record of zero-variance items dropped and the resultant changes to descriptive statistics.

Table 2: First-to-Second Eigenvalue (Percent) Drop: By Dataset. Descending sorts on DROP value. Based on smoothed tetrachoric correlations and principal components analysis with squared multiple correlations on the diagonal. EVP01=first eigenvalue in percent of all eigenvalues. EVP02=second eigenvalue in percent of all eigenvalues. 'DROP'=EVP01 divided by EVP02 -- an index proposed by Hattie, (1984, 1985).

----- Alderson Sri L. Pt. 1 ---------

GROUP	EVP01	EVP02	DROP
whole	27.69	5.90	4.69
ran sub	15.91	4.18	3.81
5th-hi	13.40	4.39	3.05
4th	14.56	5.03	2.89
3rd	13.25	6.62	2.00
2nd	8.96	6.92	1.29
1st-low	7.33	6.62	1.11

-------- Cambridge Paper 1 ----------

GROUP	EVP01	EVP02	DROP
whole	85.59	7.24	11.82
ran sub	73.44	7.08	10.37
1st-low	65.42	9.81	6.67
2nd	54.41	14.85	3.66
3rd	48.12	13.39	3.59
4th	51.35	15.32	3.35
5th-hi	47.00	17.78	2.64

------- Fall 1987 UCLA ESLPE --------

GROUP	EVP01	EVP02	DROP
whole	46.80	4.65	10.06
ran sub	30.91	3.82	8.09
1st-low	15.55	4.61	3.37
2nd	14.57	4.63	3.15
3rd	13.58	4.97	2.73
4th	13.98	5.18	2.70
5th-hi	12.73	5.45	2.34

------- Fall 1992 UCLA ESLPE --------

GROUP	EVP01	EVP02	DROP
whole	51.03	6.64	7.69
ran sub	33.47	6.30	5.31
5th-hi	23.21	7.63	3.04
3rd	19.79	7.34	2.70
4th	19.01	7.50	2.53
2nd	18.79	7.92	2.37
1st-low	13.08	9.69	1.35

------ Jones Sou. Lit. Ver. 1 -------

GROUP	EVP01	EVP02	DROP
whole	73.70	5.74	12.84
ran sub	60.93	7.46	8.17
1st-low	54.66	7.52	7.27
2nd	46.12	11.18	4.13
5th-hi	39.13	11.79	3.32
3rd	39.05	13.32	2.93
4th	37.26	13.99	2.66

------ Jones Sou. Lit. Ver. 2 -------

GROUP	EVP01	EVP02	DROP
whole	74.48	6.58	11.32
ran sub	61.04	6.55	9.32
1st-low	52.10	11.95	4.36
2nd	43.62	10.07	4.33
3rd	39.18	10.10	3.88
4th	32.86	11.12	2.96
5th-hi	28.25	13.24	2.13

------------ 1985 TOEFL ------------

GROUP	EVP01	EVP02	DROP
whole	64.67	9.13	7.08
ran sub	53.20	7.56	7.04
3rd	34.57	7.63	4.53
2nd	33.53	7.80	4.30
4th	30.45	7.78	3.91
5th-hi	23.54	7.47	3.15
1st-low	30.79	10.87	2.83

Table 3: First-to-Second Eigenvalue (Percent) Drop: Across Datasets. Descending sort on EV01/EV02 DROP. Based on smoothed tetra. correlations and principal components analysis with squared multiple correlations on the diagonal. EVP01=first eigenvalue in percent of all eigenvalues. EVP02=second eigenvalue in percent of all eigenvalues. 'DROP'=EVP01 divided by EVP02 -- an index proposed by Hattie, (1984, 1985).

RANK	GROUP	EVP01	EVP02	DROP
1.	Jones' Lit. Ver. 1: whole	73.70	5.74	12.84
2.	Cambridge Paper 1: whole	85.59	7.24	11.82
3.	Jones' Lit. Ver. 2: whole	74.48	6.58	11.32
4.	Cambridge Paper 1: ran sub	73.44	7.08	10.37
5.	1987 UCLA ESLPE: whole	46.80	4.65	10.06
6.	Jones' Lit. Ver. 2: ran sub	61.04	6.55	9.32
7.	Jones' Lit. Ver. 1: ran sub	60.93	7.46	8.17
8.	1987 UCLA ESLPE: ran sub	30.91	3.82	8.09
9.	1992 UCLA ESLPE: whole	51.03	6.64	7.69
10.	Jones' Lit. Ver. 1: 1st-low	54.66	7.52	7.27
11.	1985 TOEFL: whole	64.67	9.13	7.08
12.	1985 TOEFL: ran sub	53.20	7.56	7.04
13.	Cambridge Paper 1: 1st-low	65.42	9.81	6.67
14.	1992 UCLA ESLPE: ran sub	33.47	6.30	5.31
15.	Alderson Sri Lanka Part 1: whole	27.69	5.90	4.69
16.	1985 TOEFL: 3rd	34.57	7.63	4.53
17.	Jones' Lit. Ver. 2: 1st-low	52.10	11.95	4.36
18.	Jones' Lit. Ver. 2: 2nd	43.62	10.07	4.33
19.	1985 TOEFL: 2nd	33.53	7.80	4.30
20.	Jones' Lit. Ver. 1: 2nd	46.12	11.18	4.13
21.	1985 TOEFL: 4th	30.45	7.78	3.91
22.	Jones' Lit. Ver. 2: 3rd	39.18	10.10	3.88
23.	Alderson Sri Lanka Part 1: ran sub	15.91	4.18	3.81
24.	Cambridge Paper 1: 2nd	54.41	14.85	3.66
25.	Cambridge Paper 1: 3rd	48.12	13.39	3.59
26.	1987 UCLA ESLPE: 1st-low	15.55	4.61	3.37
27.	Cambridge Paper 1: 4th	51.35	15.32	3.35
28.	Jones' Lit. Ver. 1: 5th-hi	39.13	11.79	3.32
29.	1987 UCLA ESLPE: 2nd	14.57	4.63	3.15
30.	1985 TOEFL: 5th-hi	23.54	7.47	3.15
31.	Alderson Sri Lanka Part 1: 5th-hi	13.40	4.39	3.05
32.	1992 UCLA ESLPE: 5th-hi	23.21	7.63	3.04
33.	Jones' Lit. Ver. 2: 4th	32.86	11.12	2.96
34.	Jones' Lit. Ver. 1: 3rd	39.05	13.32	2.93
35.	Alderson Sri Lanka Part 1: 4th	14.56	5.03	2.89
36.	1985 TOEFL: 1st-low	30.79	10.87	2.83
37.	1987 UCLA ESLPE: 3rd	13.58	4.97	2.73
38.	1987 UCLA ESLPE: 4th	13.98	5.18	2.70
39.	1992 UCLA ESLPE: 3rd	19.79	7.34	2.70
40.	Jones' Lit. Ver. 1: 4th	37.26	13.99	2.66
41.	Cambridge Paper 1: 5th-hi	47.00	17.78	2.64
42.	1992 UCLA ESLPE: 4th	19.01	7.50	2.53
43.	1992 UCLA ESLPE: 2nd	18.79	7.92	2.37
44.	1987 UCLA ESLPE: 5th-hi	12.73	5.45	2.34
45.	Jones' Lit. Ver. 2: 5th-hi	28.25	13.24	2.13
46.	Alderson Sri Lanka Part 1: 3rd	13.25	6.62	2.00
47.	1992 UCLA ESLPE: 1st-low	13.08	9.69	1.35
48.	Alderson Sri Lanka Part 1: 2nd	8.96	6.92	1.29
49.	Alderson Sri Lanka Part 1: 1st-low	7.33	6.62	1.11

Figure 1: Scree Plot of Eigenvalues: Jones' Southam Literacy Test, Version 1: Whole Group

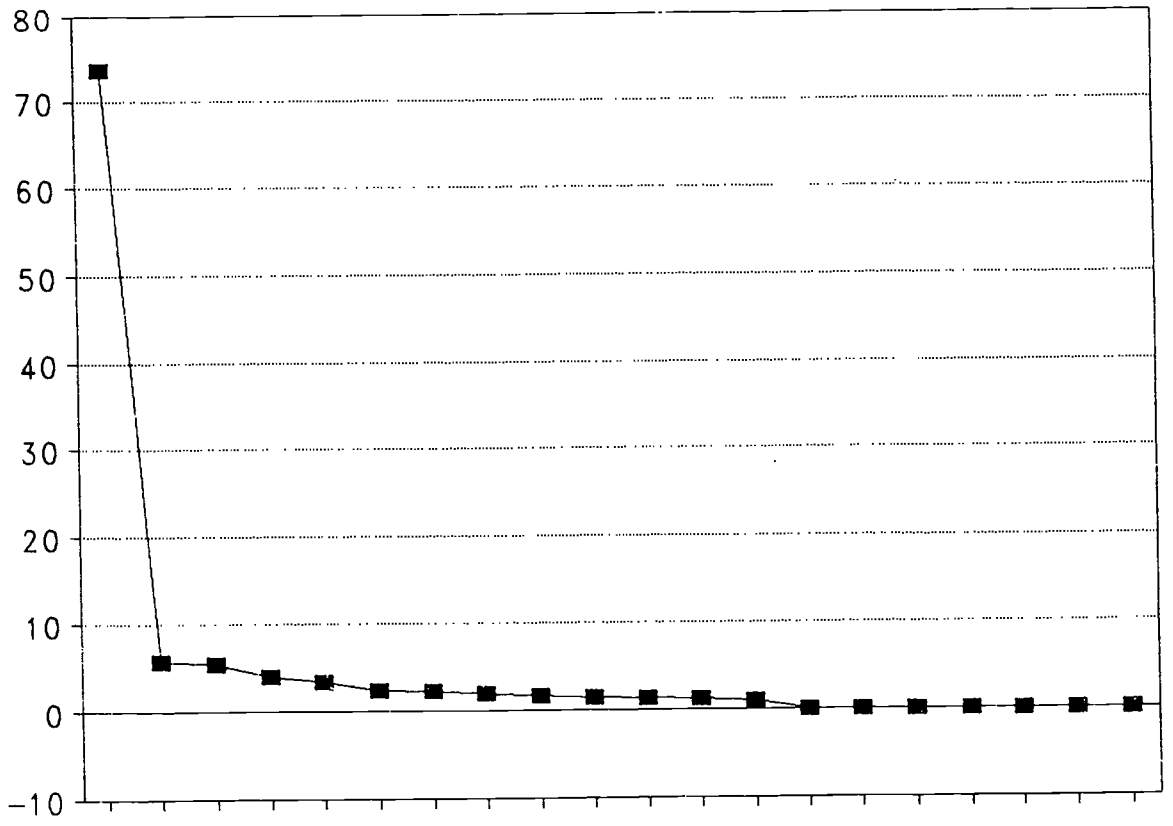

Figure 2: Scree Plot of Eigenvalues: Alderson Sri Lanka Exam, Part 1: 3rd (Middle) Ability Subgroup.

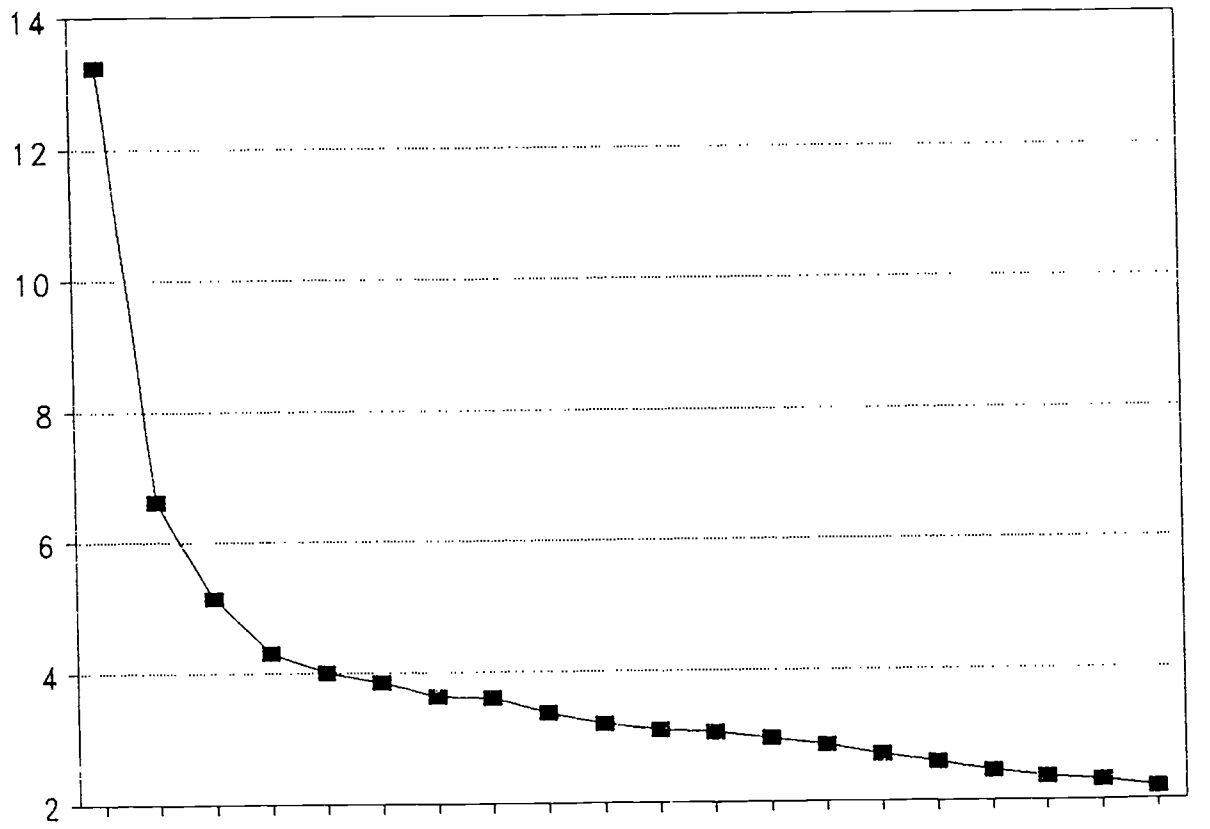

Figure 3: Scree Plot of Eigenvalues: Fall 1992 UCLA ESLPE, 1st (Low) Ability Subgroup.

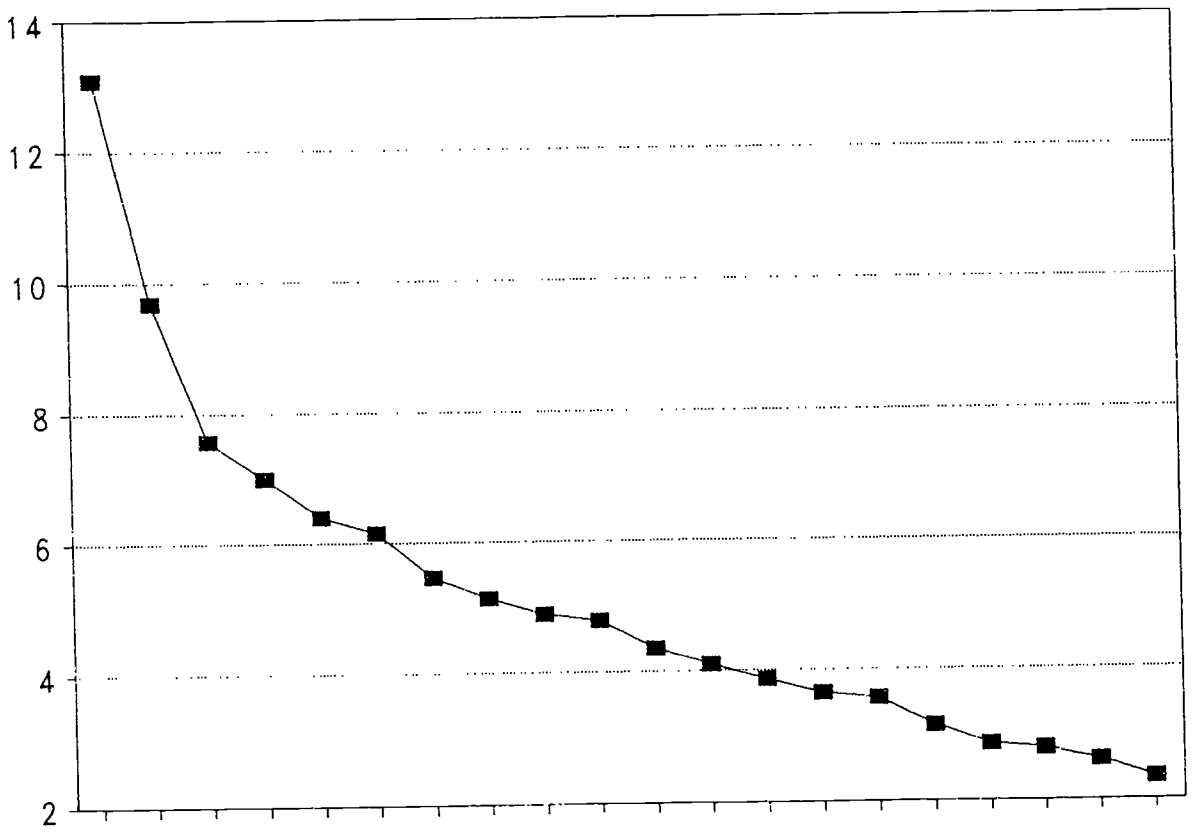

Figure 4: Scree Plot of Eigenvalues: Cambridge Paper 1, 1st (Low) Ability Subgroup.