AN EXPLORATORY STUDY OF THE HYPOTHESIS OF DIVISIBLE VERSUS UNITARY COMPETENCE IN SECOND LANGUAGE PROFICIENCY

by

ROSS PATRICK BARBOUR

B.A., The University Of British Columbia, 1968

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

English Education

Department of Language Education

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 1983

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of English Education

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date Sept. 13, 1983

Abstract

In this research Oller's question 'Is language proficiency divisible into components?' was explored by determining which of three models best fit the experimental data: a model postulating numerous specific sources of variance (the extreme divisible model), a model postulating a single, large source of variance (the unitary model), or a model postulating a large general factor and several smaller specific factors.

Following analysis of data gathered in a preliminary study, four tests which had clearly recognizable contrasts in content (grammar vs. vocabulary) and mode (listening vs. reading) were constructed to identify linguistic and method variance in a correlation matrix of language proficiency variables. These four measures were pilot tested, revised, and administered in conjunction with eight other language measures to a group of beginning-level ESL learners. The data were factor analyzed using image analysis to explore the relative congruency of the three models to the data. In addition, the relationships between the tests and the demographic variables age, sex, length of time in English Canada, and first language were also investigated.

In the factor analysis, both of the methods used to determine the number of factors to be retained in the final solution indicated three. (The methods used were the Kaiser-Guttman criterion of selecting factors with eigen values greater than one in a principal components analysis and the inspection of a varimax rotation of a full image analysis to

determine the first factor with negligible coefficients.) When transformed using a Harris-Kaiser oblique transformation (Independent Clusters), the data presented evidence for a grammar factor, a vocabulary factor and an age-related factor which may be linked closely to the hearing ability of the students. In addition, the analyses suggested the possibility that a listening-mode factor and what I have termed a 'speed of processing factor' were also influencing the variables. The factors, however, were highly correlated, suggesting the presence of a strong general factor underlying all of the measures.

The analyses of the specific relationships between each of four demographic variables (age, sex, first language, and the length of time the subject had been in Canada) and each of the twelve language variables revealed a strong negative correlation between the language measures and two of the demographic variables, age and length of time in Canada. In addition, this set of analyses revealed that the Chinese as a group performed differently than non-Chinese as a group. The analysis of sex produced no significant findings.

The conclusion of the study was that the language proficiency data in this study was best modelled by a large general factor and two specific, content-related factors, grammar and vocabulary. The possibility of specific factors related to mode was not ruled out.

Table of Contents

List of Tables

Acknowledgement

I would like to express my sincere thanks to all those people who contributed to this work. In particular I would like to thank Dr. Jamie Patrie who suggested that I attend the First Annual TESOL Summer Institute, Dr. Bernard Mohan who supported my application, and Dr. John W. Oller, Jr. who treated me to the most inspiring six weeks of my academic career. I would also like to thank Graham Evans, Geoffrey Flack, Tracy Johnson, Sherie Kaplan, Deborah Messenberg, Lucille Milligan, Donna McGee and Margaret Thompson for donating class-time for the administration of the various tests and Dr. Todd Rodgers for his valuable advice on statistical matters. I must especially thank Dr. J. Belanger who kept on saying "This looks good, but why don't you recast...?" Finally, I wish to thank Atsuko and Ian who put up with three years of "I'm off to the computer centre."

I. INTRODUCTION TO THE STUDY

1.1 Background

In the last fifteen years, there has been some controversy over the appropriate model to use in the construction of language proficiency tests. Several theorists (cf., Cooper, 1968; J.B. Carroll, 1968; B.J. Carroll, 1980; Canale and Swain, 1980) have proposed complex models of proficiency which divide language ability according to skills components such as reading, writing, and speaking, and linguistic or socio-linguistic components such as grammar, phonology and register. However, until recently, little research had been done to validate these components. Harris (1968) asked:

> What evidence do we really have, for example, to justify the neat division of most language tests into listening-speaking-writing-grammar components as the most accurate and efficient means of evaluating language 'competence'? (p. 44)

Oller (1976a, 1976b, 1979a), questioning not only the validity of the components but also the validity of the division of language competence, outlined three hypotheses which he felt would have to be supported or refuted prior to real validation of any particular components. Simply stated (see Chapter II for further explication), the alternative hypotheses were

H1) Language proficiency is divisible into unrelated components (the model of separate traits).

H2) Language proficiency is not divisible into unrelated components (the extreme unitary model).

H3) Language proficiency is partially divisible (the model postulating a large general factor accompanied by several specific factors).

Initially, Oller (1979b) felt that the third choice would be the most 'parsimonious' model (i.e. the simplest model capable of explaining the most variance). However, as a result of initial research on the hypotheses and his own pragmatic[1] approach to language, he began to advocate the second hypothesis, the extreme unitary trait model.

Very recently, though, as a result of reassessment of some of the earlier research and as a result of research by Bachman and Palmer (1981, 1982), he has renounced the extreme unitary trait model (Oller, 1981a). Bachman and Palmer (1981) say:

> As Oller (1981, forthcoming) has indicated, there now seems to be a consensus among researchers that models including both general and specific factors will provide the best explanations for language test data. (p. 450)

As yet, however, no trait other than the general trait has received strong construct validation through repetition of research on different samples of subjects. In fact, successful replication of research on different samples will be difficult since, as Powers (1982) points out, it is likely that the number and nature of underlying factors found in any set of data will depend to a significant extent on such non-linguistic variables

---

[1] The term 'pragmatic' here refers to the study of relationships between expressions in a formal system and things external to the system. (Oller, 1978. See also Ingram, 1978)

as native language, level of proficiency of the group, and, of course, the content of the tests. Thus, despite the reported consensus that H3 (the hypothesis of a large general factor and several smaller specific factors) will provide the best explanation of language test data, research investigating such data must, for the time being, begin with consideration of all three hypotheses and the implicit call for empirical support of theory.

1.2 Overview Of Experimental Procedures

The purpose of my research was to explore Oller's divisibility hypotheses concerning language proficiency (listed above) and the concomitant problem of the construct validity of certain components of that proficiency. More particularly, the present study consisted of a search for evidence of underlying factors or relevant, interpretable sources of variance in language proficiency measures. Following analysis of data gathered in a preliminary study, four tests which had clearly recognizable characteristics of content (specifically grammar in contrast with vocabulary) and mode (listening in contrast with reading) were constructed to identify linguistic and method related sources of common variance in a correlation matrix of language proficiency variables. These four measures were pilot tested, revised, and subsequently administered in conjunction with eight other language measures to a group of beginning-level ESL[2] learners at a local college. The data which were gathered

[2] English as a second language.

were factor analysed to explore the validity of the constructs of mode and content as underlying sources of variance. In addition, the relationship between the tests and the demographic variables age, sex, length of time in English speaking Canada (Lot) and first language was also investigated.

The underlying relationships among the language measures and demographic varibles of age and Lot were investigated using correlation and factor analysis. The effects of sex and first language (broadly defined as Chinese or not Chinese) on the language measures were investigated using differences of means and their associated significance levels. The study of first language was extended by deriving factor solutions for the two sub-groups, Chinese and non-Chinese, as well as for the entire sample.

1.3 Definition Of Terms

1. 'Vocabulary' is used to designate the tests with items in which the linguistic relationship among the stem, the correct choice, and the distractors is one of word meaning rather than syntax, phonology or orthography.

2. 'Structure' is used to designate tests with items in which the linguistic relationship between the stem, the correct answer and the distractors is one of syntax rather than word meaning, phonology, or orthography.

3. 'Listening-mode' refers to the fact that the test was presented entirely on tape with no printed component (other than numbers or letters to indicate choices).

4. 'Reading-mode' refers to the fact that the test used

printed materials only with no aural component involved during the administration.

5. Throughout this paper short labels are used to designate the variables in the research. Following is a brief explanation of these labels. Full descriptions of the tests are given in Chapter III, and examples are found in Appendices A to K.

A. Comp -- a composition task. The students write a short narrative.

B. Readstru -- a reading mode, multiple-choice test that focuses on structure.

C. Liststru -- a listening mode, multiple-choice test that focuses on structure.

D. Readvoc -- a reading mode, multiple-choice test that focuses on vocabulary.

E. Listvoc -- a listening mode, multiple-choice test that focuses on vocabulary.

F. Concom -- a conversation completion task. Students fill in the blanks in a dialog.

G. Oral -- an oral interview.

H. Errcorr10 -- an error correction task that has 10 items.

I. Listcomp -- a listening comprehension test.

J. Listbow -- a listening mode test that was developed from a format proposed by Donald Bowan (1975).

K. Listphon -- a listening mode, multiple-choice test that focuses on phoneme discrimination.

L. Errcorr20 -- an error correction task that has 20 items.

6. 'Progress assessment battery' refers to a set of tests given to the students at the end of each term. The format and content of these tests are changed from administration to

administration. The battery is used to help determine who is ready for the next level in the language program and who appears to need more work at the present level.

1.4 Questions And Areas Of Exploration

The questions in this study fall into three categories. First is the central problem of the divisibility of language proficiency. As I will explain in the following section, this is the problem of determining how many of the statistical factors generated by a factor analysis solution are of theoretical importance. Second is the problem of interpretation. Seven questions are presented which will aid in interpreting the factors in terms of the content and mode of the salient variables. The third category contains four subsidiary areas of exploration regarding the relations between the four demographic variables (age, length of time in Canada, sex, and first language) and the tests.

1.4.1 Divisibility

The 'divisibility problem' as presented by Oller (1976a, 1976b, 1979a) is the question whether or not language proficiency is better modelled as a single trait, the sum of several independent traits, or the sum of a single dominant trait and several subsidiary traits. The focus of his hypotheses, when restated in operational terms, is on the common or shared variance of language tests and whether there is a single general factor, several factors, or a large general factor and several minor ones. The implication is that there is

a correct choice among the three hypotheses. Cattell (1958) offers a different psychometric viewpoint of factors and their influence on manifest variables:

> In certain real but special cases it may be that a quite restricted, finite number of factors are actually operative, defining the characteristics of a population species, over the particular set of variables chosen; but in general, in our interacting universe, an infinite number of factors can and do influence the given objects and their dimensions, though the variance from a majority of the factors would be extremely small. (p. 802)

By taking this point of view in the present present research, I have altered the focus of the question from "Is language divisible and, if so, how many factors are there?" to "How many of the myriad underlying factors are important?" or, as Hakstian and Muller state it:

> The number of factors problem in this case, reduces to the task of identifying those factors whose influence is great upon the variables sampled from the domain of interest and those whose influence while real is slight. (1973, p. 461)

Therefore, the operational form of the primary question I have asked in this research is

> 1. Using the factor analytic strategies and techniques outlined by Hakstian and Bay (1973), what is the number of factors that should be retained when analyzing the correlation matrix of language and demographic variables?

1.4.2 Interpretation

In a solution which indicates that there is more than one factor significantly influencing the variables, a problem as important as the number of factors is their interpretation. Interpretation involves looking at the patterns of high and low coefficients on the rotated factors and relating them to theoretically important aspects of the variables. In this research, the important aspects of the tests are the content and the mode. The following questions are central to interpretation in this research:

1. Do the tests designated as vocabulary tests (Readvoc and Listvoc) primarily load on the same factor?

2. Do the tests designated as structure tests (Readstru, Errcorr10, and Errcorr20) primarily load on the same factor?

3. Do the group of tests designated as structure tests and the group of tests designated as vocabulary tests cluster on different factors?

4. Do all of the reading-mode tests load on a single factor?

5. Do all of the listening-mode tests load on a single factor?

6. Do the listening-mode tests and the reading-mode tests cluster on distinct factors?

7. Do the intuitively complex (in terms of content) tests (Comp, Oral, and Concom) show factoral complexity by loading on more than one factor?

1.4.3 Subsidiary Questions

During the development and administration of the tests used for the investigation of the divisibility hypotheses, I became concerned about whether or not certain characteristics of the subjects in the sample were associated with scores on specific measures. Therefore, I also explored the relationships between test scores and the four demographic variables: age, sex, first language, and the length of time the subjects had been in Canada. Although each of the questions was motivated by concerns regarding particular influences in specific tests, these initial problems are not amenable to hypothesis testing because of the post hoc nature of the analysis. Thus, I asked the more general question: Does it seem likely that these demographic variables influenced the factor analysis of the data? The aim of this section is not to test particular hypotheses but to extend the exploration of the main area of investigation to other areas which could be fruitful for further research. The following are the specific concerns and the more general questions that were generated as a result of these.

1. One of the processes commonly associated with aging is a loss of hearing. The importance of this fact in the context of this study is the possibility that the listening tests could cluster together as a result of differences in hearing ability rather than because they are measuring a 'listening-skill' dimension. That is, what might be construed as a listening-skill factor, might in reality be a hearing factor. If age can be considered as an indirect measure of hearing and if hearing

loss is the only age-related factor influencing the variables, then the expected pattern in the data would be for age to have negative correlations associated with the listening tests and zero or close to zero correlations with the paper and pencil tests. In order to investigate this in the larger context of a possible overall age-related factor, the correlations of age with the language variables were inspected, and factor solutions with and without age were compared.[3]

2. Possible bias against female subjects prompted examination of the effect of sex on the variables. The items in the two vocabulary tests were composed by a male, and I felt that this might have resulted in a predominance of words that were more familiar to male subjects than to female subjects. To explore this problem, I calculated the means on the tests for each group and the associated t-test of the difference of means and compared results for the vocabulary tests with the results for the rest of the data.

3. The large proportion (70 per cent) of Chinese speakers in the sample and in the population at large (i.e., students in the beginning level at the community college) suggested the need

---

[3] The initial motivation for investigating age as an influencing variable was my concern that older students, because of a natural loss of hearing, were being discriminated against in the listening-mode tests. It has been my experience that a number of students, particularly older ones, exhibit behavior which could be associated with a partial loss of hearing. A simple example is the tendency of some older students not to repeat high frequency /s/ and /z/ sounds at the end of words when presented with the words orally. When the words are presented in written form, the same students have no trouble producing the correct sound.

to examine the relationship between first language and performance on the tests. Two possible problems might be associated with such an unbalanced sample. The first problem is that second language acquisition may be intrinsically related to first language. Swinton and Powers (1980) suggest that different factor solutions (both in number of factors and interpretation) are associated with different language groups. If this were true, then combining the large homogenous group of Chinese speakers with the more heterogenous remainder would obscure rather than clarify the factor solution.

The second problem results from the possible cumulative effects of methods used in constructing and revising several of the measures. Listvoc, Readvoc, and Liststru are composed of items that have been tested and revised in successive administrations to samples from the same general population as the research itself. (Liststru in particular was composed of items that had undergone several revisions.) Since the majority of the subjects in the preliminary study, the pilot study, and the main research were Chinese, I was concerned that these three tests could have developed a bias against[4] that group. If this were true, and if the bias were the only factor influencing this

[4] In multiple-choice test construction distractors are chosen for their effectiveness in leading poorer students away from the correct choice. One measure of effectiveness is the number of students who choose the particular distractor. Since it is possible that some distractors are more effective against some language groups than others, those that are effective against the Chinese would be chosen rather than those which aren't simply because of the larger proportion of Chinese in the samples.

data, then the means for the Chinese speakers would be lower than the means for the non-Chinese speakers on these three tests but not on the others.

The possibility of different factor solutions for the groups is investigated by computing solutions for the full data set and for the two separate groups, Chinese and non-Chinese and then comparing the results. In addition, the possibility of differing group proficiencies is investigated by comparing the means of the two groups on the twelve language measures.[5]

4. The final auxiliary question examines the relationship between time in Canada and language proficiency. This question was prompted largely by my classroom experience. It often seems that, even within the same class, students who have been in Canada longer can carry on more extensive, 'deeper' conversations than those who have not been here so long. Furthermore, it has been my impression that this depth derives from a larger store of functional vocabulary rather than from a better grasp of language structure. This impression is supported by Powers' (1982) interpretation of the results of the research on the TOEFL (Test of English as a Foreign Language) test that he and Swinton did (Swinton and Powers, 1980). He suggests that results show:

---

[5] The appropriate method of analysis of the above two problems would involve multivariate tests of hypotheses regarding means and variance-covariance matrices such as those suggested by Kendall (1980) and Krishnaiah and Lee (1980). However, the missing data and the uneven sample sizes make such an approach technically complex, perhaps impossible and certainly beyond the scope of this research.

> ...that vocabulary, more than any other component, develops with experience or exposure. (p. 334)

To explore this, the correlations between the language measures and the reported length of time a student had been in Canada were calculated and compared.

1.5 Assumptions

Certain assumptions are made about the conditions in the research. These are

1. Students did not pass on information regarding the tests to students in the following classes.
2. Students were serious in their attempt on each test.
3. Missing data is determined by random causes.

1.6 Limitations

The sample in this research seems representative of, but not formally generalizable to, adult ESL students at the beginner level in community college classes in the Vancouver area. The formal generalizablity of the specific results of this research must be qualified by parameters from two broad areas: the demographic characteristics of the sample and the content and presentation of the curriculum.

The importance of demographic parameters in influencing the results of investigations of the divisibility hypotheses has been established (Powers, 1982; Swinton and Powers, 1980). In the sample used for the present research, the two salient demographic features that will limit generalizability are the predominance of a single language group (70 percent were Chinese

speaking) and a broad age range. Results, therefore, may not be generalizable to more heterogeneous language groups, to homogenous non-Chinese language groups, or to groups with a narrow age range.

The curriculum of the program has a specific grammatical outline and a more general subject area outline (Thompson, 1978). That is, while all instructors cover the same grammatical points, the context (and thus vocabulary) in which they are presented is more varied. The focus of the present research is on the contrast between a grammar factor and vocabulary factor and it may be that the program-wide uniformity in structure content combined with the program-wide diversity in vocabulary content will produce distinctions that would not be found in groups involved in other methods of formal instruction.

1.7 Significance Of Study

On the level of theory and construction of language proficiency models, this study will add to the body of knowledge associated with Oller's three divisibility hypotheses in four ways. First, it will add to statistical information that could support or question the psychological validity of the established linguistic distinction between the constructs of grammar and vocabulary, and the pedagogically accepted distinction between the skills of listening and reading. Second, this study has the potential to provide a contribution to future research. If strong evidence is found to support a specific (rather than a general) construct, and if the construct is reliably measured by any of the variables in the research,

then this study could supply a marker variable[6] to be used in subsequent research. Third, if only weak evidence is found, then this study can help the design of new research by indicating which content areas should be studied further and which test formats are most likely to become more effective through revision. Fourth, this study intoduces non-linguistic variables into the correlation matrix used for factor analysis. If these variables prove useful in clarifying relationships between linguistic variables, future reseachers will be able to design experiments that can control for these sources of variance.

1.8 Organization Of The Study

The basis, procedures and results of this study are presented in six chapters.

Chapter I. An introduction to and overview of the study. This chapter presents the problem, a summary of the background, the questions and areas of exploration, the assumptions, limitations and a statement of the significance of the study.

Chapter II. A review of the related research. This chapter briefly outlines several language testing models, the divisibility hypotheses and related empirical research, and the relation of the statistical tool used, factor analysis, to the process of validation of theory.

---

[6] A marker variable is a test or device which is accepted as a measure of a particular construct. These variables are used to link together research in an area by providing established reference points for new research.

Chapter III. The experimental procedures. This chapter describes the sample, the language measures, and the demographic variables. It also outlines the procedures used in compiling and preparing the raw data and the statistical methods used in the study.

Chapter IV. Preliminary analysis and pilot study. This chapter outlines the results of a preliminary analysis and a pilot study both of which were used to design the present study.

Chapter V. Summary of the findings. This chapter presents a summary of the results of the factor analyses and the exploratory consideration of correlations and differences in means.

Chapter VI. Conclusions and implications for further research. This chapter presents an over-all review of the study, the conclusions I have drawn from the results, and some suggestions and implications for further research.

II. BACKGROUND

Models that propose a variety of domains or components of language proficiency (J.B. Carroll, 1968; Cooper, 1968; Canale and Swain, 1980; B.J. Carroll, 1980) provide theoretical surveys on which to base tests. To support the validity of any particular model, there must be evidence that the components are valid constructs. Oller's formulation of three hypotheses (1979) concerning the apportionment of variance in a battery of language tests represented a call for the empirical research that he and others had recognized as being sparse (Harris, 1968; Upshur, 1976; Ingram, 1979). However, the problem of disentangling and identifying the numerous sources of variance that influence language test performance extends beyond language proficiency to the tests themselves and to the subjects who take the tests. Results so far indicate that not only may test format (Farhady, 1979) and test method (Bachman and Palmer, 1981, 1982) contaminate research results but that sample and individual variables such as first language (Swinton and Powers, 1980), first language proficiency (Johansson, 1973), and intelligence (Flahive, 1980) may do so as well. A further source of complexity is the diversity of methods of factor analysis techniques commonly used in the analysis of test battery data.[7] An investigation of Oller's hypotheses, then,

[7] One textbook on the subject (Harman, 1976) discusses nine different methods of obtaining an initial solution and twelve methods of transforming these in order to obtain better interpretivity.

entails consideration of the theoretical models of second language testing, the effect of the actual instruments on the results, the characteristics of samples used in research, and the methods used to analyze the data.

The problems of what to test and how to test when measuring second language proficiency have been addressed by many theorists in the field of second language education (cf., Harris, 1968; Oller, 1976a, 1979a; Allen and Davies, 1979; Canale and Swain, 1980; Davies, 1968). In general, the specifications that are drawn up state explicitly several separate skills and components, and the implication is that it is important to take samples from each of these areas independently in order to obtain a complete profile of the learner's language proficiency(ies). Most, if not all, textbooks on second language testing make divisions according to skills (reading, writing, speaking, listening) and some aspect of language itself such as syntax, morphology or semantics (see for example Allen and Davies, 1979; Harris, 1968; Vallette, 1977). However, the nature of the skills or components often differs depending on whether the viewpoint of the theoretical model of language that underlies the test is psycholinguistic, sociolinguistic, pragmatic, functional-notional or otherwise. As Davies (1968) points out when discussing the validity of a second language test:

> It is the test constructor's assumptions in language learning that are really being analyzed. A good test is a device for framing these assumptions....(p. 10)

J.B. Carroll (1968) focuses on language as behaviour and stresses the necessity of sampling from broad classes of stimuli and responses. He makes a distinction between productive and receptive skills and gives examples of such areas "...in which individual differences are to be sought or measured...." (p. 51) as lexicon, grammar, and phonology. In keeping with his behaviouristic approach, he presents an extensive taxonomy of possible responses to various tasks as an example of how to cover the domains of interest.

Cooper (1968), drawing from sociolinguistic theory, adds a third dimension to the usual two-dimensional, skills-by-language-component matrix used by many when proposing specifications for a test. Along one axis are the usual categories of skills (reading, auditory comprehension etc.). Along a second are placed the commonly found divisions of language aspect (morphology, syntax, etc.). As an extension of this second axis, which he labels 'Knowledge,' the concept of context is added. Finally, he proposes a third axis or dimension, Language Variety, which then provides "...84 logically distinct 'cubes' each formed by the combination of a skill, a variety, and a type of linguistic or communicative knowledge" (p. 64).

Recently two other complex models for testing language proficiency (in particular communicative competence) have emerged. Canale and Swain (1980) have proposed a theoretical framework which differentiates numerous aspects of communicative competence. They outline three general areas to be tested:

grammatical competence, sociolinguistic competence, and strategic competence. Each general area also contains subcategories such as rules of morphology, rules of syntax, sociocultural rules and rules of discourse. Each one of these would need to be sampled in a testing situation. B.J. Carroll (1980), drawing extensively on work by Munby (1978), outlines a variety of functional parameters which must be considered when drawing up the specifications for test content. These include purpose, setting, interaction, dialect and units of meaning.

Although these models may have intuitive and logical power, there has not been any strong empirical support for one over another. In the mid 1970's, Oller began to question this lack of empirical support for the validity of the various components in the different models. Aiming at the logical precedent for such models, he proposed two hypotheses regarding the divisibility of language (Oller,1976b). In hypothesis one, he suggests that language is amenable to division into unique cells defined by various skills and their intersection with linguistically posited components such as syntactic, semantic, phonological or communicative competencies. His second hypothesis is that the opposite is true: language proficiency is not divisible into subcomponents or skill areas. Later, in response to Upshur's (1976) suggestion of a logically possible middle ground, Oller (1979a) expanded the number of hypotheses to three. His third hypothesis states that language ability is partially divisible, with a large part of it taken up by a central core (perhaps to be called global proficiency), but also

includes unique skills or components which account for some of the differences between people. Placing the hypotheses in the context of expected results from language tests which attempt to measure unique abilities in supposed components or skill areas, he summarizes the hypotheses as follows:

> The Divisibility Hypothesis (H1): there will be reliable variance shared by tests that assess the same component, skill, aspect, or element of language proficiency, but essentially no common variance across tests of different components, skills, aspects, or elements:
>
> The Indivisibility Hypothesis (H2): there will be reliable variance shared by all of the tests and essentially no unique variance shared by tests that purport to measure a particular skill, component, or aspect of language profiency:
>
> The Partial Divisibility Hypothesis (H3): there will be a large chunk of reliable variance shared by all of the tests, plus small amounts of reliable variance shared by only some of the tests. (1979, p. 426)

Despite the explicit description of separate components in such models as those of Carroll or Cooper summarized above and the strong implied endorsement of H1, many theorists appear to have accepted the existence of some major latent trait that may contribute holistically to language proficiency--that is, tacit support for H3. For example, Cooper (1968) does not insist that his "cubes" define operationally distinct or orthogonal constructs. He suggests that some may co-vary. Carroll (1968), whose model Cooper elaborated upon, acknowledges the existence of such a trait and links it to the strong Verbal factor found in many factor analytic studies done on various mental

measurement batteries.

Oller, however, tended to support the unitary trait model (H2). Much of the research on the hypotheses focussed on factor analytic studies of batteries of tests given to groups of foreign students in the United States. In many cases, these studies appeared to find a single general factor which accounted for most if not all of the reliable variance in such diverse instruments as achievement tests, intelligence tests, and a variety of English language tests. (Scholz et al., 1980; Flahive, 1980; Hendricks et al., 1980; Oller and Hinofotis, 1980; Scholz and Scholz, 1979; Stump, 1978; Strieff, 1978.) Oller's interpretation of those results led him to a rejection of H1 and H3 and consequently of the type of tests which purport to test minute components of English (often called discrete point tests). He turned instead to the more integrative types of test which recognize "...the pointlessness of attempting to isolate the components of phonology, morphology, phrase structure, transformational rules, semantics and pragmatics." (1979, p. 25)

Recently, however, as a result of both reexamination of design flaws in earlier studies and the emergence of new research, Oller has changed his point of view. He (Oller, 1981) now suggests that a method of factoring based on the classical factor model rather than the method of principal components

---

[8] For a discussion of the difference see Harman, 1976, Chapter Two. For a brief explanation of the particular consequences of using the one instead of the other see Oller, 1981a.

should have been used for previous analyses.[8] In addition to this weakening of the empirical grounds on which Oller based his earlier position, recent research has presented strong evidence which supports H3. Bachman and Palmer (1981), using confirmatory factor analysis (Joreskog, 1969, 1978), have shown that, for their data, a model which includes a general factor, a reading factor and a separate speaking factor is statisically superior to the unitary model favoured by Oller. In response to these findings, Oller (1981a) states:

> The research of Bachman and Palmer has eliminated the strong version of the unitary factor hypothesis. The position that I took in several earlier publications in regard to the possibility that such a factor might prove to be the best explanation for pragmatic language processing tasks in general (Oller 1978; Oller and Hinofotis; Oller 1979, Appendix), has been proven wrong. (p. 141)

Rejection of H2 and acceptance of H3 does not simplify the search for a generalized language proficiency model. It complicates the problem by establishing the large number of hypothesized language components and skills as legitimate objects of research. As Oller (1981a) points out:

> What Bachman and Palmer have succeeded in showing is that there are undoubtedly significant factors in language proficiency tests beyond the well-established general factor. The number and exact nature of those additional factors, however, remains largely obscure.(p. 130)

Judging from the various findings of previous related research, the number and nature of factors found in future

research will depend on what is looked at, who is looked at, and how the data are analyzed. In the research that prompted Oller to reconsider his position, Bachman and Palmer (1981) produced evidence to support a speaking and a reading factor. In a separate piece of research using a similar design, they (Bachman and Palmer, 1982) found evidence of two correlated trait factors which they labelled 'grammatical/pragmatic competence' and 'sociolinguistic competence' and two method factors, writing and interview which were uncorrelated. In a study of alternate items for the TOEFL test, Pike (1979) reports finding three groupings or clusterings of scores: listening comprehension, English structure, and writing ability.

A study by Swinton and Powers (1980), which also supports the hypothesis of divisible language proficiency, introduces the complication that the results of an analysis may be closely related to the characteristics of the subjects in a sample. The study is impressive because of its size and design, and interesting because of its diverse results. The researchers factor analyzed[9] the 149-item correlation matrices derived from data obtained in the administration of the TOFEL test to seven language groups (African, Arabic, Chinese, Farsi, Germanic, Japanese, Spanish). The samples contained from 600 to 998 subjects. The different solutions established that at least three factors were necessary in each solution. Furthermore, all solutions supported the concept of a separate listening factor

[9] They used a MinRes initial solution. See Harman, 1976.

and to some extent there was agreement that a vocabulary factor was present. However, the number and interpretation of the other factors tended to differ depending on the particular language group being analyzed. On the basis of the differing means between the language groups, Swinton and Powers linked these factor differences to overall proficiency rather than first language. They proposed that:

> One hypothesis that could be investigated is the extent to which separate factors (or components of variation) are more likely to emerge as the overall language proficiency of the sample increases.(p. 15)

Such a proposal raises the possibility that a general model of language proficiency would need to be dynamically conditioned, rather than statically defined, over the range of second language proficiency.

In addition to sources of variance associated with mode, linguistic component, and possibly first language or general proficiency, a variety of other non-linguistic links to performance on second language tests have been found. Flahive (1980) found strong positive correlations between several language proficiency tests and scores on a non-verbal intelligence test (Raven's progressive matrices). Johansson (1973) found a correlation between first language performance and second language performance. Gardner (1982) reported that affective variables measured in the Attitude/Motivation Test Battery predicted (median correlation .37) French grades. The Swinton and Powers (1980) study cited earlier found a positive correlation between the factor defined as 'vocabulary' and the

two variables age and undergraduate vs. graduate matriculation status. Powers (1982) indicated that this suggested

> ...that this (vocabulary) dimension of variance was both reliably determined and distinct from the other factors. (p. 333)

The importance of these non-linguistic sources of variation in a set of data that will be analyzed is that they may, if ignored, obscure true relations between linguistic variables or create spurious ones.

As noted earlier, Oller (1981a) indicates that the method of factor analysis used in any particular research will also affect the nature of the solution and its interpretation. Generally, research that has addressed the validation of components of language has used factor analysis (Swinton and Powers, 1980; Scholz et al., 1980; Oller and Hinofotis, 1980; Bachman and Palmer, 1981, 1982). Factor analysis, however, is a general treatment covering a variety of statistical techniques which are used to discover an underlying factoral composition of a data set. A factor analysis usually follows a sequence. First, an initial solution is derived which establishes some estimate of the dimensions of the factoral space of the data. That is, it provides a general idea of the number of important underlying common factors acting in the data set. The studies by Oller and Hinofotis (1980), Scholz et al. (1980) and Hendricks et al. (1980) used the method of principal components which was subsequently criticized. Swinton and Powers (1980) and Pike (1979) used a minimum residuals method while Bachman

and Palmer (1981, 1982) relied on a maximum likelihood initial solution. Both of these latter methods are based on the classical factor model (see Harman, 1976).

Often these initial solutions are, as Hakstian and Bay (1973) suggest, "interpretively useless" (p. 29) since they do not clearly indicate the relationships between the factors and the variables. In order to provide some meaning to the factors, the axes of the space must be shifted while the projections of the variables remain stable in their relation to one another. This is referred to as a transformation (Hakstian and Bay 1973). This sequence of an initial solution followed by a transformation is often repeated with different numbers of factors before a preferred solution is found. Which type of transformation is eventually chosen will depend on the relative simplicity of the structure of the solution and whether the solution is meant to be exploratory or confirmatory.

Harman (1976) has pointed out that "...a given matrix of correlations can be factored in an infinite number of ways." (p .4) The important points in choosing a preferred solution are, he says, statistical simplicity[10] and scientific meaningfulness. Or, as stated by Hakstian and Bay (1973):

> The guiding principle in such transformation is Thurstone's notion of simple structure, or the idea that each factor should be interpretable in terms of (or have high loading by) a small number of variables, with the remaining variables relatively free of the influence of (or loading near zero on) that factor.

---

[10] Harman gives a more detailed outline of the concept of simple structure. (1976, p. 97-98)

(p. 29-30)

To further aid in the choice of a solution, Hakstian[11] has divided factor analytic research into:

> that motivated by either taxonomic or explanatory interests on the part of the investigator...
>
> The taxonomic view of factor analysis regards factors as merely convenient groupings or clusters of variables -- groupings that carry little construct validity or epistemological status.
>
> The explanatory view of factors and factor analysis, on the other hand, regards factors as causal agents -- valid and replicable constructs that determine the covariation among the phenotypic constructs in the domain of interest. (p. 16)

Research which concerns the construct validity of various components of language usually takes the 'explanatory' view of factors.

Factor analysis, both taxonomic and explanatory, can also be divided into 'exploratory' and 'confirmatory.' Philosophically, the difference is whether the researcher has a hypothetical structure in mind when he approaches the data and wishes to confirm this or not. Statistically, the two are different in that in an exploratory analysis "...the shape of the final solution is not influenced by conditions outside of the analysis" (Hakstian and Bay, 1973, p. 69). In confirmatory factor analysis, on the other hand, constraints are put on the solution by setting a target matrix which embodies a theoretical model proposed by the researcher. Joreskog (1969, 1978)

---

[11] Hakstian and Bay, 1973. See also Hakstian and Muller, 1973.

outlines one method of determining the statistical significance of the closeness of fit of such a solution. That there is a difference between the philosophical and statistical understanding of the term 'confirmatory' is important in the context of language research. Bachman and Palmer (1981) tested "...over 20 different causal models...." (p. 78) against their data. That is, they were using a statistical confirmatory analytic technique in an exploratory manner[12] not 'confirming' the validity of a pre-existing theory.

In summary, the results of research on the divisibility hypotheses and the problem of the validation of language proficiency constructs are characterized by four themes. First, there are the various theories (linguistic, sociolinguistic and psycholinguistic) of competence and performance. These, of course, determine the nature of the actual measuring devices from which springs the second problem: the identification of method, format, or mode variance in the results. Third, there is the nature of the sample itself. In the multi-dimensional, real universe that the subjects bring with them to the testing environment, a variety of psychological, experiential, and demographic variables have been identified that seem to have significant effects on performance on language tests. Finally, there are the methods used in analyzing the data. Different approaches have been shown to bring different results and

---

[12] This is not meant to be criticism of their work. Joreskog (1978) points out that in some cases exploratory techniques may actually obscure particular types of structures within the data.

interpretations (see in particular Oller, 1981); as yet no single method can be said to have the unqualified support of all in the field. Research which seeks to discover the nature of language proficiency and whether it is unitary, divisible, or dominated by a global trait but also composed of subsidiary, specific traits must address all four of these themes.

III. DESIGN AND PROCEDURES

3.1 Population

The subjects in the study were E.S.L. students in a metropolitan community college. They were adults (18 years and older) from a variety of linguistic, cultural and educational backgrounds. Analysis of the demographic variables shows 14 different languages (see Table I), and an approximately even (55 percent male: 45 percent female) division of sexes (see Table II). The age ranged from 19 to 73 years old.

Table I - Breakdown of sample by first language

	Language	Frequency
1.	Chinese	121
2.	Vietnamese	22
3.	Japanese	4
4.	Punjabi	4
5.	Spanish	4
6.	Gujarati	3
7.	Greek	2
8.	Korean	2
9.	Portuguese	2
10.	Hindi	1
11.	Italian	1
12.	Polish	1
13.	Russian	1
14.	Tagalog	1
15.	(non-chinese)	9*
	Missing	2

(* These cases were known to be non-Chinese but the actual language was not known)

Typical previous groups included a wide range of backgrounds: farm workers with little or no education and professionals such as doctors, dentists, or engineers. The majority of the students are immigrants or Canadian citizens. People who have come to Canada on student or visitor's visas are not permitted to enrol in this program. Some of the subjects may hold diplomatic visas.

Table II - Breakdown of cases by sex

MALE		FEMALE	
N	%	N	%
93	(55)	76	(45)

Missing cases: 12 (not reported)

The subjects' ability in English can best be illustrated with an outline of the hierarchy of the entire program. The college program has three levels: beginners, intermediate and advanced. In each level there are two sub-levels, lower and upper. Those students who wish to go on to study in content areas in colleges or universities generally have to take another year of language studies beyond the 'advanced' level. In summary then, there are six sub-levels or steps leading from zero proficiency through to a level before college preparation. The data were gathered from students at the second sub-level

(Upper Beginners). Several of the tests in the analysis were those in an assessment battery used to promote students from Upper Beginners to Lower Intermediate.

3.2 The Tests And Demographic Variables

Twelve language proficiency measures and four demographic variables provided the data for the investigation. In the descriptions and explanations which follow, I have grouped the language proficiency measures into four categories. First are four measures (Liststru, Readvoc, Listvoc, and Readstru) which were included in the research to mark contrasts in mode (listening vs. reading) and content (grammar vs. vocabulary). Second are four subtests (Concom, Errcorr10, Listcomp, and Oral) from the progress assessment battery administered to the population at the end of the term. Third is the composition score which is from data collected in an internal college project on the development of a composition rating scale. In the fourth category are three supplementary tests (Listphon, Listbow, and Errcorr20).

3.2.1 Tests Developed For Interpretive Purposes

Table III presents summary statistics from four tests that were constructed specifically to identify contrasts in mode (listening and reading) and content (grammar and vocabulary) in the interpretation of the results of the factor analysis.[13] These tests all show moderate reliabilities, ranging from .65 to

[13] The development and pilot testing of these four measures is outlined in Chapter IV, Preliminary Study and Pilot.

.74. Low to moderate internal reliabilities may derive either from error variance or from the response of the variable to more than one underlying source of variance (Magnussan, 1967). As Borg and Gall point out (1978), error variance in measures will obscure finer distinctions that would otherwise be made apparent. Thus, while being a drawback in that they may indicate error variance which is clouding real distinctions among the variables, the low to moderate values of the reliability coefficients will not invalidate any distinctions that are found. If, on the other hand, the reliability estimates have been depressed by factoral complexity of the variables, then this will be revealed in the

Table III - Summary statistics for tests used for interpretive purposes

	Mean	s.d.	No. of Items	N	Rel
LISTSTRU	12.3	3.7	28	155	.66
READVOC	16.0	4.1	27	146	.69
LISTVOC	8.0	3.2	20	167	.65
READSTRUC	18.9	3.9	30	164	.74

analysis because the variable will load on more than one factor. Among the tests in Table III, Listvoc is the only test that shows the adverse effect of being too difficult. Since the mean (8.0) is only about one standard deviation (3.2) above the chance score of five, there is probably some error variance

being generated by guessing. As noted earlier, this effect would be reflected in the reliability.

1. Liststru (listening-structure) is a multiple-choice English grammar (structure) test in listening mode. (See Appendix A for script of introduction, sample items, and sample answer sheet.) It is an extended and revised version of the multiple-choice listening structure test used in the pilot study and in content is almost identical to Readstru described below. The prototype of Liststru was simply a reading-mode (paper and pencil) grammar test transformed completely into a listening-mode test with all parts, stem and options, being heard by the subject. This test was included to investigate the effect of mode. If listening mode is a unique source of variance (different from both content and reading mode) then this test should exhibit factoral complexity. That is, there should be common variance with Readstru and with some other factor that could be identified as strongly related to the mode of listening.

2. Readvoc (reading-vocabulary) is a multiple-choice vocabulary test in reading mode. (See Appendix B.) It was designed to identify the presence, if any, of a "vocabulary" factor underlying the twelve variables. The format and mode are identical to Readstru. Therefore, if format and mode are sources of variance, this test will overlap to some degree with Readstru, even if there is a component of language proficiency that could be labelled 'vocabulary.' The extent of the overlap will give some indication of the strength of mode and format in

contrast to content as sources of variance.

3. Listvoc (listening-vocabulary) is the aural form of Readvoc (see Appendix C). That is, it is a multiple-choice English vocabulary test in listening mode. In fact, as pointed out in Chapter IV, Listvoc and Readvoc are merely presentations in different modes of items randomly selected from the same item pool. The test was included to highlight and make interpretable a vocabulary factor if one could be educed. It forms a clear mode/content contrast with Readstru.

4. Readstru (reading-structure) is a multiple-choice English grammar test in reading mode. (See Appendix D.) Like Liststru, it was included in order to identify a grammar or structure factor if one was influencing the set of variables. This test is not a version of the reading-mode structure test described in Chapter IV although it is very similar. It is one module of a multiple choice English grammar test that was being developed at the college at the time of the research.

3.2.2 Subtests From The Progress Assessment Battery

Four of the measures used in the research were the four subtests in the progress assessment battery given to students at the end of the term. Table IV presents the summary statistics for these tests. The reliability of Errcorr10 (.75) and Listcomp (.70) are moderate. Relative to the length of the test, the reliability (.75) of the ten item Errcorr10 is very high. According to the Spearman-Browm formula for correcton for attenuation (Ebel, 1974) the reliability of this test would be

.90 if made the same length (30 items) as Readstru. The reliability is no doubt helped by the test's independence from error variance created by guessing.

Table IV - Summary statistics of subtests in the assessment battery

	Mean	s.d.	No. of Items	N	Rel
CONCOM	7.3	1.9	(10)	181	-
ERRCORR10	5.9	2.4	10	181	.75
LISTCOMP	12.8	3.5	20	181	.70
ORAL	14.4	2.6	(25)	181	-

The two tests (Oral and Concom) which were subjectively graded have no measure of reliability from which to estimate error variance. The narrow standard deviation (1.91) of Concom suggests that the test was not making as clear distinctions among students as the other measures and consequently low correlations between this test and any others may be as much a reflection of this as of a difference in language dimension. In addition, this test showed a slight ceiling effect with a third of the students obtaining 90 percent or greater. Both the narrow standard deviation and the ceiling prevent it from displaying an accurate representation of the relationship between this kind of task and content and the others in the analysis. Interpretation of the results of the analysis are

tempered by this information.

The oral test does a better job of spreading out students than does Concom. In addition, there is no ceiling effect on the distribution so these two problems will not be present in the interpretation of the factor analyses or correlations.

1. The test labelled Concom (conversation completion) is a completion type of exercise in which the student writes the answer in a blank (see Appendix E). In this particular type of test the student reads a short introduction which outlines a situation. This is followed by an incomplete dialog in which several of the sentences are replaced by blank lines. The student's task is to fill these blanks with appropriate, grammatically correct (though not necessarily complete) responses.

This test was graded by the students' own instructors who used the following guidelines for marking. Each blank was assigned an equal percentage of the total. The written responses were first considered for appropriateness. If the response did not follow from or lead into the rest of the dialog, it was given zero for that part. For example in the following

a: And how are you today?

b: _____

a: Oh, that is too bad. How long have you felt like that?

if the student wrote "Fine, and you?" then the mark for that

response was zero. Similarly, if the response was not comprehensible because of structure, word usage, spelling, or handwriting, the mark was zero. Those parts which did not receive zero were checked for grammatical accuracy and spelling. A single point was removed for each major structural error (incorrect deletion of a verb or subject, wrong tense, word order etc); half points were removed for spelling errors or minor structural errors (deletion or insertion of articles, plural or third person 's', countable nouns treated as uncountable and vice versa). A student's score was the sum of the remaining points.

On the face of it, the combination of the task and the evaluation method clearly lead to complexity, covering reading comprehension, situational proficiency, structure, vocabulary, and spelling. The score was included in the analysis to see if such hypothetical complexity would be borne out statistically. Unlike the compositions, this exercise has no measure of inter-rater reliability. It was not estimated in the assessment battery procedure, and the original product of the student was not available afterwards for re-evaluation. I felt that if the measure displayed low communality with the other variables and had an erratic or unstable behavior in the analysis then it could be dropped. Experience with a similar tests in the pilot and previous administrations of the battery suggested that it would show moderate communality with the other tests (r=.37 to .60. See Table V.)

Table V - Subtest-test correlations of Concom with previous assessment batteries

Date		R	N
Aug	1979	.60	89
Feb	1980	.59	73
May	1980	.68	98
June	1980	.47	119*
Aug	1980	.44	108*
Oct	1980	.37	131*

(Those tests marked with * are correlated only with the reading/writing subtest total)

2. Errcorr10 (error-correction, 10 items) is the second test in the assessment battery. It is an error correction format with 10 items (see Appendix F). In this format the student is given a short reading passage of fifty to one hundred words. Parts of the passage (words or phrases) are underlined. The students' task is to determine whether the underlined portion is in error or not and correct it if it is. While such tests may contain a variety of errors (vocabulary, usage, spelling, or structure), this test included only structural errors. This format had been used extensively on the different assessment batteries at the beginner levels over the preceding four years. In each analysis the format had shown good reliability (.69 to .80) and a good tendency to spread students out despite the short length. (See Chapter IV for statistics on two other such tests.) With the items focussing as they do on

structural errors, the test can be labelled a grammar test and was included in the analysis with the expectation that it would show common variance with Readstru.

3. The test labelled Listcomp (listening-comprehension) is a multiple choice form of the type of test commonly termed listening comprehension (see Appendix G). In this form of test, subjects hear a conversation between two people, three to five lines long and lasting four to fourteen seconds. (In this particular test they heard the conversation twice.) Afterwards, the students hear several questions. These questions ask for recall of details of the conversation or for inferences about the people and their location or activities. Following each question, answer choices are given. The student circles the letter on the answer sheet which corresponds to the correct choice. (In this form of the test for the present research, neither the conversations nor the questions and choices were presented in print.)

This general format had been included on the assessment battery four times in the three years preceding the research. In all four administrations the results were unsatisfactory because of low reliabilities. Despite this, inspection of the individual item statistics suggested that it was possible to create an effective test using this format but that care would have to taken to avoid making one that was too difficult. Because of this history, I also felt that a twenty-item test might be more effective and reliable than the usual ten-item one. However, including an extended version of this type of

test in the assessment battery was impractical because of time constraints. Instead, an entirely new test was created using new items modelled on those from the previous tests which displayed satisfactory item characteristics. I intended to administer one test before the assessment battery and one within that battery, then combine the two scores to produce a single listening comprehension mark. When the first module was administered, it was obvious that it was still far too difficult for the target population. The assessment-battery module was simplified and lengthened, and the number of options decreased from four to three. Of these, only two of the options were aural. The third option was that neither of the first two was correct.

4. The Oral test is an eight-minute, four-part, guided interview with an instructor. (See Appendix H for guidelines and sample score sheet.) It consisted of a warmup, a free-speaking period, a question-making section and a language-use section. In each part, the focus of the interview was different and the students' responses were evaluated according to slightly different criteria. The weighting of the parts was approximately equal, but the criteria did tend to emphasize accuracy of structure.

The oral interview method of testing language proficiency has a great deal of face validity. However, as with the conversation completion exercise (Concom), this measure has no estimate of inter-rater reliability. The practical problems associated with obtaining such estimates are large. Mullins

(1980) has shown that the best reliability can be obtained when interviewers are given a general scale on which to base their judgments. In addition, previous analyses of the battery showed significant positive correlations of similar oral assessments with the total test (see Table VI). These ratings were given under similar conditions to those in this study: raters were not the students' own instructors and had no knowledge of the students' previous performance on any language tests.

Table VI - Correlations of previous oral assessments with progress assessment batteries

Date	R	N
Aug. 1979	.69	89
May 1980	.88	98
June 1980	.70*	119
Dec 1980	.77	73

(* does not include oral test in total test)

If the matrix of language variables does allow a multi-factor solution, the oral measurement will have important interpretive power because it can not be logically associated with reading mode or paper-and-pencil tests as such. Thus it has the important potential of distinguishing linguistically related common variance from mode or method related common variance.

3.2.3 Composition

The composition score (Comp) is based on data gathered in a separate project done by a committee at the college to develop a program-wide scale for rating student compositions. The final form of the scale consisted of descriptions of five skill levels in three hypothesized components of writing skill: semantics, syntax and orthography (See Appendix I). In the initial stage of development, samples of student writing from several levels were scrutinized and descriptions of the written work at the various levels were composed. These descriptions were distributed to other instructors for suggestions on clarification of wording. Next, a group-training session was held during which instructors used the scale to evaluate samples of students' work. In the final stage, each Wednesday for three consecutive weeks each student in the ESL program wrote a composition based on a set of pictures which depicted a storyline involving several people. They had one hour to complete their work. After each writing, the papers were graded separately, first by the students' instructors and then by another instructor. Because the results of these evaluations would be used to promote the students, if there was a difference of three or more points between the first two raters the paper was evaluated by a third rater and the discrepant grade was eliminated.[14] At the end of the three weeks, then, three sets of two (or three) ratings on the students' composition

[14] This is in agreement with the procedure recommended by Diederich (1974).

proficiency were available.

For the purpose of the present research, only ratings on the final compositions were used. (See Table VII for summary.) This was done for two reasons. First, I felt that the instuctors had become more adept at using the scale by that time and that their evaluations had become stable and more in agreement with the scale. This was indicated by the increase in the inter-rater correlation (Pearson product-moment) from .73 on the first set to .89 on the final set. In addition, the final composition was written in the same time period as the rest of the measurements done for the research.

Table VII - Summary of ratings and computed score used for composition grade

	Mean	s.d.	Total Possible	N	Rel.
First Rater	7.36	2.46	25	175	-
Second Rater	7.40	2.46	25	175	-
Composition	14.8	4.78	50	175	.89(a)

(a) correlation of first and second rater

The score used in the research was the sum of the scores given by the two raters or in the sixteen cases where a third rater was required, the two scores closest together. In no case did the third rater fall exactly between the first two scores.

3.2.4 Supplementary Tests

In both the preliminary research and the pilot study, only two listening-mode tests were included. However, Harman (1976) suggests it is necessary to have at least three tests loading on a factor to define it and Gorsuch (1974) recommends five. Consequently, following the pilot, two more listening-mode tests were developed and pilot tested. An additional error correction format test was also administered. Table VIII presents the summary statistics for these three supplementary tests.

Table VIII - Summary statistics for three supplementary tests

	Mean	s.d.	Number of Items	N	Rel.
Listphon	46.3	9.3	62	134	.92
Listbow	15.1	4.6	29	157	.80
Errcorr20	7.7	3.71	20	153	.81

1. The test labelled Listphon (listening-phoneme) is a listening test in which the students must distinguish between vowel phonemes (see Appendix J). It is an extension of a commonly-used classroom sound discrimination exercise. In such an exercise, the student hears three words and is required to determine which one (first, second, third) is different from the other two. For this test, two other options were included: the choice of all words different or all words the same. The

material for the test was taken from a pronunciation book containing minimal pairs (Nilson and Nilson 1973). Two each of thirty-one of the thirty-three vowel contrasts given in the book were chosen.[15]

This phoneme discrimination test was pre-tested on a sample of thirty-nine students. The results of this pre-testing showed a reliability of .92 and a good spread (sd= 8.25). Once its reliability was determined, no item analysis or revision was done despite there being numerous items that were obviously ineffective. I felt that since certain languages have more trouble with some contrasts than others, the elimination of "ineffective" items might bias the particular test strongly against the Chinese, who made up 70 percent of the population.

The test was created and included because it was short (thirteen minutes), easy to create, and represented an extreme end of the discrete-integrative test item scale. I felt that if a listening factor were found and if this particular measure were closely related to it, then it could be used effectively as an indirect measure of that factor.

2. The test labelled Listbow (listening-Bowen) was a listening test based on a format developed by Bowen (1975) which he called an integrative test of English Grammar (see Appendix K). He suggested that it:

> ... measures the ability of a subject to reconstruct

---

[15] Two contrasts based on the phoneme /ɔ/ (as pronounced in 'caught' in some American dialects) were omitted because this distinction is not made in Canadian English.

> obscured words by means of sentence analysis, carried out not as a separate academic task, but in the normal procedure of understanding what the sentence says. It is a task built on the assumption that the ability to handle reduced redundancy is a valid measure of linguistic competence. The reduced redundancy in this type of test is a consequence not of deliberate deletions or masking by superimposed noise, but of the reductions, assimilations, and contractions that normally accompany sentence production by native speakers functioning in a relaxed, informal context. (p.2)

The test requires a student to listen to a sentence and write the second word. Usually this word has been reduced by a contraction or run together with the preceding or following word. For example the student might hear "Where'd he go?" and be expected to write 'did'.

The script as presented by Bowen was too long for the present study. In his research, Bowen used sixty different items which he presented to the subjects twice each in the same sitting. First, all items in which the focal reduction resulted in the same sound were grouped together and then following that the same items were presented again in random order. As a result the total test was 120 items long.

For a pilot run of this format, the first half of the script as presented by Bowen was recorded and administered to an Upper Intermediate class. The results of this pilot indicated that the test was too difficult for the Upper Beginners level and consequently would be inefficient. The format was kept but fifteen pairs of simpler items aimed at the approximate level of the target population were created. The items in each pair had as their focal reduction the same sort of syntactic

relationship. (For example a subject pronoun in a simple 'be' question.) To ensure a good range of marks, the first fifteen items were spoken at a reduced speed and the second fifteen at a speed approaching natural speech. A pilot run of this test suggested it would be satisfactory (r=.80, sd=3.82) and after some minor revision, a new tape was made and administered to the subjects.

To ensure consistency in the results, I marked the papers. As mentioned earlier, the instructions to the students stipulated that only the second word in each sentence be written down. In a few cases the students wrote down more than one word for several of the items. Where this happened, the items were marked as incorrect, even if the correct word was included. In three cases (of 160) the students did this for all of the attempted items. The Listbow tests for these subjects were deleted completely. During the administration of this form it was found that the focal reduction of one item[16] was indistinguishable even to native speakers. This item was not included in any of the subsequent analyses.

3. The final supplementary test is Errcorr20 (error correction, twenty items) which has exactly the same format as Errcorr10 (see Appendix F) except that there are twenty underlined items. During the design of the experiment, I was not sure whether there would be an error correction type of

---

[16] This was item twenty. The reduction of the pronoun 'her' in the sentence "Is this her book?" was interpretable as either 'her' or 'your.'

exercise on the assessment battery. Since this format had proved efficient in the pilot study and in previous assessment batteries, I felt it imperative to include such a test (Errcorr20) in the research. When I found that there would also be one on the battery (Errcorr10, I kept both.

3.2.5 The Demographic Variables

Students reported data used for the demographic variables on a form which they filled out on the day following the progress assessment battery. First language, age, and sex were used as reported by the students. To reduce potential errors in arithmetical calculations, students were asked for the year and month they arrived in Canada. Data for the variable length of time in Canada (Lot) were calculated from this information.

3.3 Administration Procedures

The tests developed for the purpose of the present research were administered in conjunction with a progress assessment battery at the end of the regular four month term of instruction. I did not supervise the administration of the assessment battery (Oral, Concom, Errcorr10, Listcomp). However, the same guidelines were followed by the instructors for both the assessment battery and the research tests. The research tests (Listbow, Liststru, Readvoc, Listvoc, Listphon, Errcorr20) and the composition were administered in the two-week period preceding the administration of the assessment battery. The administration of the battery and Readstru was done in one day during the regular class period. The administration of the

oral test was spread over two days preceding the administration of the assessment battery.

In the administration of each test, the instructors followed the same general rules: no dictionaries or notes were allowed; no help was given by the instuctor after the actual test had started. For the paper and pencil tests, the instructor waited for all students to be in the room and then presented the examples and reviewed them with the students. If the students had problems, the instructor continued to give help with understanding the task. Generally, the kinds of tasks used (choosing a correct answer, writing in a word or sentence, or writing a connected set of sentences) were all common class exercises and presented nothing novel to the students. In the listening tests, some of the tasks were unfamiliar and consequently more examples were given for these tests. The Listbow test, perhaps the most novel, included a total of eight examples. For all these listening-mode tests the students had the option of hearing the introduction with the examples several times.

At the end of each of the six tests constructed and administered specifically for the present research, the answer sheets and test papers were collected by the instructor and given to me. In order to gain the cooperation of the instructors, I allowed them to use the test papers the following day for teaching purposes. For the evening classes, this was allowed on the same day, as there was no security problem. In addition, as far as was possible, the answer sheets were graded

for the instructors and a class list of results handed back to them the following day.

On the day following the progress assessment test, students filled out a form which requested a variety of biographical details including those four (age, sex, first language, and length of time in Canada) used in the present research.

The administration of the oral test was conducted over a period two days by four instructors who were experienced in the use of the method. Students left the regular classroom, went to an office for the interview, then returned to class and sent another student.

3.4 Statistical Procedures

The statistical procedures can be divided into three areas: data preparation and description, the factor analyses, and the subsidiary analyses. Prior to analysis, the raw data had to be transcribed into a form that could be read by computer. This was done using a microcomputer for data entry and for transfer of the data to disk storage at the University of British Columbia. When this was completed, the computer at the University of British Columbia was used to obtain summary statistics, do the factor analyses and complete the subsidiary analyses.

3.4.1 Data Preparation And Description

After the administration of all of the tests, the answer sheets, the progress assessment battery booklets, and copies of the Readstru and composition scores were collected, collated and placed in class groups. Next, the data was transcribed onto magnetic disc. This was done using a microcomputer for data entry. Because one of the steps in the research was to calculate reliablity estimates, it was necessary to encode the option chosen by each student on each item of the multiple choice tests (excluding Readstru).[17] In order to handle the resulting 40 thousand discrete pieces of data effectively and to diminish chances of entry error, I wrote a data entry program for a microcomputer. This program was designed so that as each set of test responses was entered, the length of the set was checked against the length of the appropriate test. If these did not match, a signal was given and that particular test was re-entered. This avoided gross errors that might have resulted from entering a set of responses under the wrong heading or from adding or deleting a single response in a test and entering the following responses displaced by one item number.

After the data had been transferred from floppy disk to storage on the computer system at the University of British Columbia, an error check was made by comparing a listing of the data with thirty randomly drawn sets of the original papers.

---

[17] This had been done on optical-read score sheets and analyzed separately.

This error check revealed negligible error rates in the three categories of data: item responses, raw test scores, and biographical information. In this thirty-subject sample, a total of seven item responses (out of 217 items for each subject) were incorrect for an error rate of .1 percent. In the same sample, two errors occurred in the entry of the thirteen biographical variables. This represents an error rate of less than .5 percent. The inspection of the eleven raw scores and subject measures for each subject in the error sample revealed two errors. Although this represents only a .6 percent error rate for entry in this subset of the variables, the nature of the particular errors found in this check prompted a check of the full data set. The two errors that were found occurred in the same two variables (Oral and Concom). The values for measures had been transposed and it appeared that the errors were systematic rather than random. Consequently, the values for these two variables were rechecked in the entire data set and three additional errors were discovered. The errors that were found in each category were corrected, of course, but the overall error rate indicated that the data could be used as entered and corrected, without further editing.

After the error check, the computing facilities at the University of British Columbia and subroutines from the Statistical Package for the Social Sciences (SPSS)[18] were used to:

---

[18] Nie et al., 1975

1. score the multiple-choice tests

2. compute alpha reliabilities for the multiple choice tests

3. transform the reported year and month of arrival in Canada to a single variable

4. assemble a 195-subject data set that included the twelve language variables and the four demographic variables.

In order to make this data base more stable, all cases which were missing data on four or more of the variables were deleted. This reduced the total number of cases to 181. Following the deletion of these cases, SPSS was used to obtain descriptive statistics and histograms for each of the variables. This information was used in evaluating the tests as language measures, to determine the suitability of all of the variables for further statistical analysis, and to supply an overview of the demographic features of the sample.

In preparation for the factor analysis, two alternative methods were used to replace missing data in the score matrix. First, the step-wise regression procedure in SPSS was used to obtain regression estimates (or 'predicted scores') of the missing data. The second approach was to replace each missing value with the mean of the respective variable. As a result of these procedures, three data bases were available for analysis: one with missing data points, one with missing data replaced by regression estimates, and one with missing data replaced by mean scores.

To decide which of the three data sets to use in the factor

analyses, the FACTOR procedure in SPSS was invoked and a principal components solution followed by a varimax rotation was performed on all three sets. The solutions showed very little difference. The mean difference of the highest and lowest loading for each variable on each factor was less than .060 and the single largest difference was .13 (on age, 3rd factor -.74, to -.87). A similar comparison was done on the twelve language variables. The mean difference of the highest and lowest loading for each variable on each factor in this set was less than .055. The single largest difference was .12 (Listvoc on the third factor-- .78 compared to .90). This similarity among the solutions for the different methods of treating missing data indicated that a choice among them could be based on criteria which were external to the actual solution. The data set which had missing values replaced by mean values was subsequently chosen for all further analyses. The greatest advantage to using mean scores to fill in missing data is that it is cheaply and easily done. This was important in the later analyses because the data was split into two groups and of course the missing had to be filled in again using the new means.

3.4.2 Factor Analysis

The principal statistical method was factor analysis which was used for three purposes. First, as mentioned above in section 3.4.1, it was used to decide the most appropriate treatment of missing data. Second, it was used to choose the best subset of variables for the final solution. Finally, it was used to arrive at representative solutions to the central

problem of the research: demonstrating the divisibility of language proficiency and giving meaning to the components.

Incomplete component analysis followed by a varimax rotation was used in the analyses done to determine which missing data treatment to use and in determing the final subset of variables. In these cases it was the comparability of clusterings of variables and agreement on the number of factors for each solution that was of interest. The Kaiser-Guttman criterion (Harman, 1976; Hakstian and Bay, 1973) of eigenvalues greater than 1.0 was used throughout these analyses to serve as a standard for determining the number of factors.

In deriving the final solutions, Hakstian and Bay's (1973) strategies for exploratory factor analysis were followed. First, an image analysis was used, followed by a varimax rotation. Then, combining inspection of the results of this with the results of a principle components-varimax combination and using the criteria recommended by Hakstian and Bay (1973), a decision was made about the number of factors. Finally, an image analysis was done again, this time retaining the number of factors that had been indicated by the earlier procedures. This was followed by a Harris-Kaiser oblique transformation (independent clusters), which allows the axes (and thus the factors) to be correlated. It also has the effect of bringing the factor solution closer to the criteria of simple structure.

These strategies, applied first to the full 181 case set of data, were repeated for the Chinese-speakers and for the non-Chinese speakers. The missing cases in the two sub-groups were

filled with group means.

All factor solutions were generated using either SPSS or the Alberta General Factor Program (Hakstian and Bay, 1973)

3.4.3 The Subsidiary Analyses

The subsidiary analyses were done using SPSS. The data were first divided according to sex, and then means and t-tests were calculated for the twelve language variables and Lot and age. Next the data were regrouped according to first language (Chinese and non-Chinese) and the means and t-tests were again calculated. One of the correlation matrices that were included in the output for the factor analyses was also used in the subsidiary analyses.

IV. PRELIMINARY STUDY AND PILOT

4.1 Preliminary Study

In the fall of 1979, I began a project of revising and standardizing an ESL progress assessment battery in the English Language Training program of a metropolitan community college. Much of this work involved gathering statistics on each of the items and subtests already being used and then using this information to improve the battery as a whole. In addition to the project of revising the established battery, I began a parallel project of experimenting with a variety of new items and tasks. I intended that this serve the dual purposes of expanding the number of usable items and subtests and of investigating Oller's three divisiblity hypotheses.

The target population of the battery and of the experimental items was a group of beginner-level adult ESL learners of mixed linguistic and educational background and ages.[19] The purpose of the battery was to determine which students were proficient enough to move on to more advanced language study and which needed to continue to work at their present level. At the time the first set of data was gathered, the students who took this battery and the experimental items were at the fourth level of a ten-level program ranging from no ESL proficiency through to pre-college entrance. The labels for each level were Beginners 1 (B1), Beginners 2 (B2), Beginners 3

---

[19] This is essentially the same population as that in the research.

(B3), Beginners 4 (B4), Intermediate 1, Intermediate 2, Intermediate 3, Intermediate 4, Lower Advanced, and Upper Advanced. Although students often took longer to move through a level, the progress battery was administered every two months.

In February 1980, as part of the validation procedure, the battery and an experimental listening test were given to both the B4 level and the level below (B3).[20] Summary statistics are presented in Table IX. These statistics allow several comments to be made concerning the validity of the battery, its subtests and the experimental listening test. First, the differences in means between the two levels, B3 and B4, show that each of the subtests was clearly distinguishing between the groups. The difference in means on the total test is particularly large, which can be taken as one demonstration of its validity. Furthermore the overall estimate of reliability for the battery was moderately good for both the target B4 level (Cronbach's alpha=.79) and for the combined levels (Cronbach's alpha=.68). Since the tests had already been inspected by a number of instructors for content and face validity, it can be said generally that the battery as it stood and was used then was a valid measurement of language proficiency.

To explore the validity of the subtests in the context of Oller's divisiblity hypotheses, I factor analyzed the data using a truncated principal components solution followed by a varimax

---

[20] Descriptions of the subtests in the battery and of the experimental listening test are given in Appendices E to H and L respectively.

Table IX - Descriptive statistics of subtests in preliminary research

TEST		B3 (N=71)	B4 (N=73)	B3&4 (N=144)
1:Listen 1	Mean	4.41	6.11	5.27
10 items	SD	1.76	1.53	1.85
	Range	0-8	2-9	0-9
	Hoyt Rel	.35	.20	.43
2:Listen 2	Mean	5.78	7.53	6.67
10 items	SD	1.7	1.83	2.09
	Range	1-9	3-10	1-10
	Hoyt Rel	.40	.58	.59
3:MC (Reading)	Mean	7.35	11.18	9.29
18 items	SD	2.33	2.81	3.21
	Range	2-13	5-18	2-18
	Hoyt Rel	.46	.50	.66
4: Error	Mean	5.52	9.12	7.35
Correction	SD	2.27	2.44	2.97
15 items	Range	0-11	3-14	0-14
	Hoyt Rel	.46	.50	.66
5:Composition	Mean	6.14	9.16	7.67
15 marks	SD	3.04	2.64	3.21
	Range	0-14	2-14	0-14
6:Conversation	Mean	4.89	8.33	6.63
Completion	SD	2.83	2.29	3.09
12 marks	Range	0-11	0-12	0-12
Oral	Mean	-	15.15	-
Interview	SD	-	4.38	-
25 marks	Range	-	5-23	-
Total Test	Mean	34.09	51.44	42.89
(Excluding	SD	7.54	9.22	11.72
Oral)	Range	18-53	23-65	18-65
	Alpha Rel	.49	.79	.68

rotation. The results are presented in Table X.

While Factor I of Table X is not readily interpretable, Factor II is clearly related to listening mode since both of the subtests in listening mode load heavily (Listening 1 =.736, Listening 2 =.761) on Factor II. The loading of .35 of the Oral test on Factor II is consistent with any language proficiency model which included listening as a component. In an interview the subject will receive aural cues for his/her responses.

Table X - Varimax rotated factor solution for seven English tests in preliminary tests, level B4 (n=73)

TEST	FACTOR I	FACTOR II
1. LISTENING 1	.009	.736
2. LISTENING 2	.129	.761
3. MULTIPLE CHOICE	.343	.386
4. ERROR CORR.	.728	.157
5. COMPOSITION	.703	.162
6. CONVER.CONPL.	.809	-.117
7. ORAL	.610	.350

Oller and Hinofotis (1980) also found some support for the notion of a listening factor. The moderately weak loadings of the Multiple Choice test on both Factor I (.343) and Factor II (.386) show that this test has low communality with the other tests and also suggest that the test may be factorally complex. (That is to say it may be measuring more than one component of language.) Inspection of the items in the Multiple Choice test produces evidence to support this conjectured complexity, as the following examples illustrate:

#6. Why _____ come to my party last week?

1. don't you
2. you don't
3. didn't you
4. you didn't

#4. "Where is Bill?"

"I saw him go outside with a hammer, a saw and some nails. I think he is going to _____ ."

1. work in the garden
2. cut the fruit tree
3. paint the garage
4. fix the fence

#12. "How was the test?"

"I got the best mark in the class."

" _____ ."

1. What a pity.
2. You have my sympathy.
3. Congratulations.
4. Better luck next time.

In number six, the correct answer is determined structurally, using past tense and word order. In number four, the correct answer is determined by the meaning of the words, connecting "hammer," "saw," and "nails" with "fix" and "fence." In number twelve, the answer is determined by the recognition of

the correct social formula. As these examples show, the Multiple Choice test contained content from three different areas: vocabulary, grammar, and social idiom.

In summary, although the preliminary study was done on a group of measures that were not specifically designed to investigate the divisibility of language proficiency, the study revealed two potentially successful avenues for investigation of Oller's hypotheses: contrasting listening mode tests with tests in other[21] modes, and contrasting content in the form of vocabulary, grammar, and socially acceptable idiom or formulas. These two avenues were the subject of a pilot study.

4.2 The Pilot Study

To investigate the implication of the preliminary study that mode and content constitute significant, contrasting sources of variance in language test data, I constructed four multiple-choice tests, two vocabulary tests and two grammar tests (see Appendices A to D), using items drawn from an item-bank that had been developed during the revision of the progress assessment battery. One of the vocabulary tests and one of the grammar tests were then converted to listening mode by tape recording the items, with stem and numbered options each repeated but not presented on paper. Subjects were given answer sheets on which they circled the number of the correct choice.

---

[21] The tests which did not load on the putative listening factor represent three other modes of testing: reading, writing, and speaking.

In August, 1981, the new reading-mode grammar test was incorporated into the regular progress assessment test as the multiple choice section. The three other experimental tests were administered in conjunction with the battery in four of the eleven Upper Beginners classes[22] and the data were gathered for analysis. Unfortunately, the quality of the sound of the listening comprehension test in the assessment battery was poor, resulting in the elimination of the test. The oral interview was not included because at that time, the college's testing policy had changed. Previously, all students took the oral test, but at the time of this study only students who scored above (about) 60 percent on the paper and pencil test (including the listening test) were allowed to take the oral.

The test statistics are presented in Table XI and the results of a truncated principal components solution followed by a varimax rotation are presented in Table XII.

In Table XII, the loadings of .83 for the reading-mode structure test and .77 for the listening-mode structure test strongly associate Factor I with the measurement of grammar. Similarly, the loadings of .78 for the reading-mode vocabulary

---

[22] It should be noted that in the interim between the preliminary research and the pilot study, the original levels B3 and B4 were merged into a single level, Upper Beginners. At the same time, the term was extended from two months to four months. The net result was that at the time of the pilot study, there were a greater number of students taking the progress assessment battery.

Table XI - Descriptive statistics of subtests in pilot study (n=60)

NAME	ITEMS	MEAN	SD	RANGE	HOYT REL
1. MC STRUC(READ)	20	11.43	3.25	6-18	.65
2. ERROR CORR.	20	8.57	4.30	1-18	.80
3. MC VOCAB(LISTEN)	16	7.45	3.13	1-14	.69
4. MC STRUC(LISTEN)	19	8.20	3.07	2-16	.58
5. MC VOCAB(READ)	27	16.20	4.81	3-25	.78
6. COMPOSITION	(15)	8.13	3.70	0-15	-
7. COMPLETION	(12)	6.85	3.12	0-12	-

test and .88 for the listening-mode vocabulary test link Factor II with the measurement of vocabulary. Although this pattern of a content-related contrast in factors is distinct from the mode-related contrast found in the preliminary study, it does support the theoretical analysis made of the complexity of the multiple choice

Table XII - Varimax rotation of PC solution for 7 subtests in the pilot study (N=60)

NAME	FACTOR I	FACTOR II
1. MC VOCAB (LISTEN)	.07	.88
2. MC VOCAB (READ)	.29	.78
3. MC STRUC (LIST)	.77	.30
4. MC STRUC (READ)	.83	.09
5. ERR.CORR	.83	.06
6. COMPOSITION	.76	.38
7. CONVERSATION COMPLETION	.80	.19

test in the preliminary study. It was very unfortunate that the listening comprehension test had to be deleted, for it would appear on the surface and from the results of the preliminary research that it is a different kind of test. However, despite the drawback, there was enough evidence to suggest that in the presence of an expanded battery of tests, these four tests (the multiple-choice, reading-mode grammar and vocabulary tests, and the multiple-choice, listening-mode grammar and vocabulary tests) would act as effective marker variables at least for a vocabulary/structure dichotomy and possibly for a mode contrast.

V. FINDINGS OF THE STUDY

The factor analyses done on the three sets of data[23] present evidence for a grammar factor, a vocabulary factor, and an age-related factor (possibly hearing). In addition, the analyses suggest the possibility that a listening-mode factor and what I have termed a 'speed of processing factor' are also influencing the variables. The analyses of the specific relationships between each of four demographic variables (age, sex, first language, and the length of time the subject had been in Canada--Lot) and each of the twelve language variables reveals a strong correlation between the language measures and two of the demographic variables, age and Lot. In addition, this set of analyses reveals that the Chinese as a group performed differently than the non-Chinese as a group. Because of these findings, age has been included in the matrices which are analyzed in the divisibility study and the Chinese and non-Chinese speakers have been treated separately as well as together. The analysis in which sex was the dependent variable produced no significant findings.

---

[23] The data were first analyzed as a complete set then divided into those who spoke Chinese and those who did not. For convenience, I refer to the groups as the combined group, the Chinese speakers, and the non-Chinese speakers.

5.1 The Factor Solutions And The Divisibility Hypotheses

For the divisibility hypotheses, the most significant feature of the three different solutions (the combined group, the Chinese speakers and the non-Chinese speakers) is that the two different statistical criteria[24] agree on a three-factor solution. That is, they both support some form of divisibility in language proficiency. Of equal importance to the divisibility hypotheses is that in all three solutions, two of the factors are characterized by having high coefficients from the language measures. In the solution for each group, one factor consistently provides evidence for the validity of a grammar or structure factor while another factor supports the concept of a vocabulary factor. The third factor in each solution was associated with high coefficients from age. However, the configuration of the other coefficients on this factor suggest three different interpretations depending on the group for which the solution was done: hearing (the combined group and the Chinese-speaking subset), listening, and 'speed of processing' (the non-Chinese speaking subset). The interpretive and theoretical power of all of these factors, though, must be tempered with the caution that in each solution the factors were correlated and thus cannot be considered as truly independent sources of variation in the data.

---

[24] These were the Kaiser-Guttman criterion of selecting factors with eigenvalues greater than one in a principal components analysis, and inspection of a varimax rotation of a full image analysis.

5.1.1 Factor Solution For Entire Set

Table XIII shows the solution arrived at for the combined group. The three matrices that are presented are the Phi matrix, which shows the correlation of the factors with each other; the pattern matrix, which gives an indication of the relative strength of each factor in the variable; and the structure[25] matrix, which gives the correlation of the variables with the factors.[26]

The first factor in Table XIII can be defined by the clustering of high coefficients from the grammar tests: Readstru, Errcorr10, and Errcorr20. The other tests with moderate coefficients on this factor (Comp, Concom) are consistent with its interpretation as a grammar factor.

---

[25] The word "structure" is a term from the field of factor analysis and is unrelated to structure in the sense of grammar.

[26] When looking at an oblique factor solution (one in which the factors have been permitted to correlate) the pattern matrix gives the clearest picture of the underlying factoral composition of the variables. This is the matrix that is used for interpretation of the factors. The Phi matrix reveals how much the factors are correlated. If the correlation between two factors is close to zero, then the factors are acting independently and represent true differences in dimensions. (When the correlation is set at zero in a solution, the solution is termed orthogonal.) As the correlations between factors increase, their interpretability as separate influences decreases and so does theoretical power. Harman, (1976) gives some examples of how to interpret oblique solutions in Chapter IV.

Table XIII - Image analysis followed by Harris Kaiser (Independent clusters) on full (n=181) data set

	PATTERN MATRIX			STRUCTURE MATRIX		
	I	II	III	I	II	III
ERRCOR10	0.74	-0.01	0.03	0.72	-0.51	0.61
READSTRU	0.74	-0.07	0.12	0.68	-0.50	0.46
ERRCOR20	0.63	-0.07	0.11	0.67	-0.46	0.59
CONCOM	0.46	-0.01	0.01	0.46	-0.32	0.39
COMP	0.38	-0.02	0.36	0.69	-0.56	0.69
LISTSTRU	0.24	-0.12	0.19	0.49	-0.44	0.47
READVOC	0.07	0.13	0.56	0.46	-0.34	0.52
LISTCOMP	0.02	-0.16	0.51	0.57	-0.56	0.65
LISTVOC	-0.28	0.06	0.77	0.33	-0.32	0.49
ORAL	-0.07	-0.37	0.31	0.45	-0.55	0.52
LISTBOW	0.00	-0.46	0.17	0.48	-0.60	0.53
LISTPHON	0.16	-0.45	-0.18	0.33	-0.43	0.30
AGE	0.15	0.71	-0.18	-0.20	0.46	-0.22

CORRELATION MATRIX OF FACTORS (PHI)

	I	II	III
I	1.00		
II	-0.71	1.00	
III	0.85	-0.76	1.00

VARIANCE OF FACTORS

I	II	III
2.00	1.10	1.49

However, the low coefficient of Liststru on this first factor weakens the interpretation. Since Errcorr10, Errcorr20, and Readstru are presented in print, this may also be a 'paper and pencil' factor. Yet for verification of this argument, the value of Readvoc, which is also a 'paper-and-pencil' test, should be closer to the value of the grammar tests (.63 to .74) rather than almost negligible (.07).

Factor II is also interpretable. Of the three tests which have moderate coefficients on this factor (Listphon, Listbow,

Oral), two are listening tests. These two tests (Listphon and Listbow) are the only two of the five listening tests in which items are not repeated. Furthermore, these two are the least contextualized of the listening tests. That is, they contain the least amount of redundant information. (Listphon has none at all.) This lack of extra information makes the items much more difficult for those subjects who, because of hearing problems, miss part or all of an item. I included Age, the variable with the highest coefficient on this factor, as a variable to be analyzed specifically because I had noted that several of the older students were hard of hearing. In this sense, I had intended it to act as an indirect measure of hearing. Consequently, a possible interpretation of this factor is that it represents the influence of the physiological variable of hearing rather than a linguistic component of the tests.

The moderate coefficient on Factor III from Oral is also consistent with its interpretation as a 'hearing' factor. Comprehension is included in the evaluation guidelines for the interview and it is conceivable that interviewers have attributed manifestations of hearing problems to indications of a weakness in English. Asking for repetition or answering the 'wrong' question would be two such behaviours.[27]

The interpretation of the third factor in Table XIII is

---

[27] Another possible explanation for the moderate coefficient of oral on this factor is that it indicates a bias on the part of the interviewers against older people. If this were the case, of course, the loading is unrelated to hearing.

straight forward. The two vocabulary tests, Listvoc and Readvoc have the highest coefficients on this factor. The fact that they are in different modes gives a great deal of strength to the interpretation of this as a vocabulary factor. Furthermore, the three other tests (Oral, Comp, and Listcomp) which have moderate-to-high coefficients on this factor are not only entirely consistent with this interpretation but also add strength to it. These three tests are in different modes which suggests strongly that the common feature in the different tests that causes them to cluster together is one of content rather than mode. Consideration of the content of these three tests also supports the interpretation of Factor III as a vocabulary factor. Correct use of words will have a positive influence on both oral and composition grades. In the listening-comprehension test, ten of the twenty questions require the students to draw inferences about the location, actions or characters involved in the short dialogs. Such inferences draw heavily on an understanding of specific words and phrases used in the dialog.

The final important aspect of this solution is the Phi matrix in Table XIII. It indicates that the three factors are highly correlated. One explanation for correlated factors is that the factors themselves are responding to a single, higher-order factor. Such an explanation in this solution lends support to the concept of the indivisibility of language proficiency. The two linguistic factors would have to be considered as different manifestations of a single global

proficiency factor. A different explanation lies in the measuring devices themselves: they can be viewed as tapping different conglomerations of several (hypothetical) components of language proficiency. For example, as noted in the discussion on the third factor, several of the tests, while not being designated as 'vocabulary' tests, can be thought of as responding to differences along some 'word knowledge' dimension in addition to their putative purposes. Composition in particular must be considered a task involving the integration of grammar and vocabulary. In fact, of the twelve language measures, only one, Listphon, does not integrate structure and vocabulary into either the task or the product. With such inherent theoretical and practical complexity, it is not surprising that some variables show statistical complexity[28] and that the factors themselves are correlated. To establish the construct validity of distinct linguistic factors, it would be necessary to overcome this inherent problem of complexity. Some methods of creating less complex tests such as structure-free vocabulary tests are discussed in Chapter VI.

In summary, the solution for the full set indicates the influence of a structure factor (Factor I), an age-related factor which I have argued is best designated as a hearing factor (Factor II), and a vocabulary or word knowledge factor (Factor III). The high correlation of the factors in this oblique solution reflects the integrated nature of most of the

[28] A complex variable is one which has moderate to high coefficients on two or more factors.

tasks and may also show that, in fact, there is only a single significant source of variance influencing the language variables.

5.1.2 The Subset Of Chinese Speakers

Table XIV presents the solution arrived at for the more homogeneous subset of Chinese speakers. Here again the three factor solution is the preferred one. The pattern of coefficients on this matrix is similar to the one on the full data-set matrix. This is not surprising, of course, since the Chinese group represents two-thirds of the combined group. Factor I is still clearly a structure or grammar factor. In this solution, the coefficient of Liststru (.45) is higher than in the solution for the combined group (.24). This adds strength to the interpretation of Factor I as a grammar factor as opposed to a reading-skill or method factor because Liststru is in listening mode whereas the other three grammar tests (Errcorr10, Errcorr20, Readstru) are in reading mode.

The age-related factor in this solution, although similar in configuration, does not account for as much variance as the same factor in the solution for the combined group. This is apparent not only through the relatively lower values for Oral, Listbow, Listphon, and age, but also through the differences in the variance of the factor in the two solutions: 1.10 (Table XIII) in the combined group and .73 (Table XIV) in the Chinese speakers. In terms of underlying influences on the linguistic variables, this information suggests that in this solution, age

Table XIV - Image analysis followed by Harris Kaiser (Independent Clusters) on Chinese subjects (n=121)

	PATTERN MATRIX			STRUCTURE MATRIX		
	I	II	III	I	II	III
ERRCOR10	0.83	-0.02	-0.10	0.76	-0.50	0.68
ERRCOR20	0.68	0.08	0.07	0.69	-0.40	0.64
READSTRU	0.68	-0.08	0.01	0.73	-0.52	0.67
LISTSTRU	0.45	-0.02	0.07	0.49	-0.31	0.47
CONCOM	0.48	-0.09	-0.14	0.42	-0.31	0.36
COMP	0.49	0.13	0.36	0.74	-0.41	0.73
LISTCOMP	0.14	-0.05	0.49	0.62	-0.44	0.65
READVOC	0.06	0.09	0.51	0.47	-0.26	0.51
LISTVOC	-0.30	0.07	0.76	0.34	-0.19	0.44
ORAL	-0.10	-0.22	0.51	0.51	-0.47	0.55
LISTBOW	-0.06	-0.33	0.37	0.49	-0.51	0.52
LISTPHON	0.29	-0.38	-0.16	0.38	-0.47	0.33
AGE	0.09	0.60	0.06	-0.24	0.50	-0.22

CORRELATION MATRIX OF FACTORS (PHI)

	I	II	III
I	1.00		
II	-0.64	1.00	
III	0.91	-0.60	1.00

VARIANCE OF FACTORS

I	II	III
2.40	0.73	1.55

is acting much less as an indirect measure of some physiological (or psychological) impediment to language learning or production than in the solution to the larger group.[29]

Characterized by salient coefficients from the two vocabulary tests, Factor III is clearly interpretable as a

[29] The average ages of the two subsets of the data were very close; 31.7 for the Chinese speakers and 32.2 for the non-Chinese. This suggests that the difference in strength of the age-related factor is not merely a reflection of an age difference in the two groups.

vocabulary factor. The difference between the third factor in this solution and the third factor in the solution for the combined group is that both Oral and Listbow have larger coefficients in this solution. As noted in the discussion on the combined group, Oral having a high coefficient on this factor is consistent with the vocabulary interpretation. It can also be argued that a moderate coefficient from Listbow is consistent with the vocabulary interpretation. Bowan describes his listening test as an integrative test of English grammar. While the focus of each item is a structural point, the path to the correct answer lies through the comprehension of the entire sentence. Quite possibly, it was word knowledge that presented the greatest barrier to comprehension. If this were the case, then the test would be functioning more as a vocabulary test than a grammar test.

The Phi matrix in this solution shows once again that the factors are highly correlated. In this solution, Factor I and III (the structure factor and the vocabulary factor) are even more highly correlated (.91) than in the solution for the combined group (.85). Such a high correlation between factors weakens the theoretical power of the factors as distinct traits despite their apparently clear identification in the pattern matrix. Whether this correlation of factors is a result of the complexity of the majority of the variables as mentioned earlier, or whether the identification of the two factors in the solutions as grammar and vocabulary factors is merely 'fooling oneself with factor analysis' as Nunnally (1978) puts it, will

have to be determined in further research. However, it is clear from the solutions on these two data sets (and from the solution on the non-Chinese set discussed in the following section) that the most fruitful direction in any research which attempts to define separate factors in language proficiency at the level of ability of subjects in this study will be to focus on grammar and vocabulary.

In summary, in the factor analysis for the Chinese group, the solution indicated three factors, one identified as a structure or grammar factor, one identified as a vocabulary factor and the third as an age-related factor, possibly related to the physiological variable of hearing.

5.1.3 The Non-Chinese Speakers

Although it is open to criticism on statistical grounds, the solution for the non-Chinese speaking group is interesting in that it is both similar to and different from the other two solutions.[30] Table XV shows that in this solution too, Factor I, with salient coefficients from the three reading-mode grammar tests, remains interpretable as a structure factor. In

[30] In a fourteen-variable matrix there are 91 different correlations. Among this many, especially with only 60 subjects, the probability is high that some values are spuriously large or small. Consequently it is difficult to know if a particular coefficient is a result of chance correlation or truly reflects the influence of an underlying factor. Second, this smaller group is in a sense a microcosmic reflection of the complete data set. Within the group of 'non-chinese speakers' are 22 Vietnamese speakers. If first language is a direct factor in learning English as a second language (Chapter VI suggests an alternative interpretation) then once again having a large homogenous group within an otherwise heterogenous sample would be expected to obscure the results.

addition, Factor III, with high coefficients from the two contrasting mode vocabulary tests, is still identifiable as a vocabulary factor.

Table XV - Image analysis followed by Harris Kaiser (Independent Clusters) on 60 non-Chinese subjects-retaining three factors

	I	II	III	I	II	III
ERRCOR20	0.72	0.08	-0.07	0.72	0.42	0.35
ERRCOR10	0.65	-0.04	0.13	0.70	0.37	0.46
READSTRU	0.67	-0.01	-0.11	0.61	0.28	0.24
CONCOM	0.54	-0.06	0.10	0.56	0.27	0.35
COMP	0.37	0.40	0.02	0.59	0.61	0.44
READVOC	0.19	-0.06	0.50	0.42	0.31	0.56
LISTVOC	-0.13	0.08	0.66	0.26	0.37	0.63
LISTCOMP	-0.04	0.47	0.31	0.38	0.63	0.63
LISTSTRU	0.14	0.39	0.12	0.41	0.53	0.41
LISTPHON	-0.03	0.62	0.33	0.13	0.43	-0.00
ORAL	-0.07	0.62	0.01	0.25	0.59	0.31
LISTBOW	0.10	0.69	-0.08	0.42	0.69	0.35
AGE	0.31	-0.65	-0.03	-0.04	-0.49	-0.21

CORRELATION MATRIX OF FACTORS (PHI)

	I	II	III
I	1.00		
II	0.53	1.00	
III	0.53	0.55	1.00

VARIANCE OF FACTORS

I	II	III
1.98	2.20	0.95

Factor II is age-related, with a coefficient of .65 for age on that factor. However, it is certainly different in nature from the age-related factor found in the larger sets. In both of the previous analyses only two tests (Listbow and Listphon) clustered with age. In this solution there are six: Comp, Listcomp, Liststru, Listphon, Oral, Listbow. Because of this, it is difficult to think of it as simply a hearing factor. On

the other hand, four of the six tests that have high or moderate coefficients on this factor are listening-mode tests and another one is the oral interview which, as has been noted, does involve listening. Thus, this solution on the non-Chinese set could be construed as supporting the concept of a 'listening-skill' factor or perhaps a bi-polar hearing/listening factor.

Another interpretation for this factor is that it measures the speed with which language processing tasks are handled. This interpretation stems from a common feature of tests of such apparent mode and content diversity as a composition task, an oral interview, Bowan's listening test and a listening comprehension test.[31] All of these tasks can be viewed as dynamic, involving a marshalling of several skills or language components under the pressure of time. Age, too, has its highest coefficient on this factor but it is negative. One trait associated with aging may be a slowing down of the speed with which language (or any other information) is processed. This would be reflected particularly in tests and measures in which the information flow was continuous and not controlled by the recipient. Listening tests and the oral interview would both fall into this category. A slowing down of language processing would also be reflected in language <u>production</u> tasks that were constrained by time such as the oral interview (again)

---

[31] This interpretation may also be relevent to the interpretation of the third factor found in the complete data set and in the Chinese subset. There, too, the salient coefficients belonged to Oral, Listcomp, Listbow, and Comp. On the other hand, in these solutions age did not cluster with these four.

and the composition task. Establishing the validity of such a trait is outside the scope of the data in this study, but some suggestions regarding its implications for further research are presented in Chapter VI.

The Phi matrix in Table XV is also of interest. The correlations between the three factors range between .53 and .55 which is somewhat less than those in the solution for the Chinese-speaker subset. This difference suggests that in this sample there is more distinction between the factors and that there was more heterogeneity in skills in the non-Chinese group than in the Chinese-speaking group.

In summary, the analysis of the non-Chinese speaking subset reiterates the presence of both a grammar and a vocabulary factor. It also introduces the possibility of a hearing/listening factor (as opposed to a strictly hearing factor found in the previous analyses) or a speed of processing factor.

5.1.4 Summary

In conclusion, the factor analyses provided moderate evidence that a multiple factor solution is preferred in the analysis of this matrix of language and demographic variables. It also produced support for the argument that knowledge of English grammar and knowledge of English vocabulary are identifiable sources of variance within the matrix. The analysis also suggested but did not clearly support the notion of a source of variance related to the modality (e.g., listening) of the instrument. In addition, I have presented a

brief description of a hypothetical 'speed of processing' factor that would be amplified in tests in listening mode though present in other modes. In this factor a major source of variation would be the speed with which language was processed and since listening tests present language in a stream uncontrolled by the subject, these tests would be particulary influenced by such a trait. In all solutions, the factors were highly correlated, leaving open the question of how many of the identified factors are truly significant.

5.2 The Demographic Variables And Their Relation To The Tests

Inspection of the means and correlations of the variables suggested that three of the four demographic variables (age, Lot, and first language but not sex) might be significantly influencing the shape and interpretation of the factor solutions. Therefore, a series of subsidiary analyses was done in which the solutions for different subsets of the data were compared. While adding age to the language-variables matrix did clarify and simplify the solution, adding Lot did not.

5.2.1 Age

The correlation of age with each test and p-values associated with a one tailed t-test are presented in Table XVI. As noted in Section 1.4.3, age (as an indirect measure of hearing) was expected to have negative correlations associated with the listening tests and zero or close to zero correlations with the paper and pencil tests. As Table XVI shows, this pattern was not found. While three listening-mode tests

(Listbow, Listphon, and Listcomp) do show strong negative correlations with age, two reading-mode tests (Readstru and Errcorr10) do also, which suggests that age may also be associated with some other negative effect on language proficiency. Furthermore, Oral, which is not a listening test, has a larger negative correlation than three of the listening tests (Listcomp, Liststru, and Listvoc). While some of this can be attributed to the fact that an oral interview has a listening/hearing component, it seems reasonable to expect that a hearing problem could be compensated for by the interactive

Table XVI - Correlations of age and length of time in Canada (LOT) with language measures

	AGE			LOT		
	R	(N)	P	R	(N)	P
LISTBOW	-.43	146	.000	-.14	138	.057
LISTPHON	-.41	126	.000	-.43	118	.000
ORAL	-.36	168	.000	-.03	157	.375
LISTCOMP	-.28	168	.000	.00	157	.481
READSTRU	-.19	151	.009	-.09	141	.140
ERRCOR10	-.18	168	.011	-.04	157	.315
LISTSTRU	-.17	142	.021	-.02	132	.402
COMP	-.11	162	.081	.03	152	.377
CONCOM	-.11	168	.078	-.10	157	.099
LISTVOC	-.11	155	.094	.21	145	.005
ERRCOR20	-.03	143	.333	-.07	133	.210
READVOC	-.00	134	.480	-.01	124	.440

nature of the task. That is, a subject being interviewed could

ask the evaluator to speak louder and more clearly.[32]

One possible explanation for the pattern that appears in Table XVI is that the variable age is acting as an indirect measure of two (or possibly more) otherwise independent influences for example, hearing and the 'speed of processing' factor proposed earlier. In tests which are affected by both of these influences, the effects would be amplified, making the statistical correlation large. Where only one or the other is acting, the correlation would be proportionately reduced. Possibly this 'speed of processing' factor in some tests (for example Listbow and Listphon) is combining with the hearing problem to increase the correlation with age, but in other tests (especially in reading and writing mode) acting alone, and thus producing a lower correlation with age. In still other tests (Listvoc or Liststru for example) this factor may not be an influence at all, leaving the correlation of the test with age the result of the influence of the hearing factor. Clear characterization of these hypothetical age-related traits will require research which includes a direct hearing measure and several tests designed to accentuate differences along the hypothesized 'speed of processing' dimension.

Whatever the underlying causes, the pattern of correlations in Table XVI is a convincing argument for including age in the factor analysis. It is quite possible that a group of variables

---

[32] As noted earlier, though, these very actions may be interpreted by the interviewer as indicating poor language comprehension.

which are not directly related to each other cluster together in a solution because of a common influence of age on the correlations. Without age to identify the cluster, the interpretation would be misleading. To investigate the possibility that an age-related factor was producing spurious linguistic clusters in the analyses, I compared several factor solutions which included and excluded age in the matrix of language variables. These solutions are presented and discussed in Appendix M. The comparison indicated that retaining age in the matrix clarified the relationships among the linguistic variables.

5.2.2 Length Of Time In Canada

Table XVI also presents the correlations of Lot with each of the language measures. Since the correlation of age with Lot was .42*(n=155,p=.000), it is difficult to know how much of the negative and near-zero correlations of Lot with the language variables is a result of the mediating effect of Age.[33] That there are negative correlations of Lot with nine of the language measures is somewhat surprising. It seems intuitively unreasonable to expect that any of the language measures would be in fact negatively related to the length of time a person had been in Canada for this could imply a loss of language proficiency over the period a subject had been in Canada. It is

[33] Two unrelated variables can show a statistical correlation if they are both correlated to a third. Similarly a relationship that does exist between two variables can be hidden if the two variables are both correlated to a third variable but in an opposite manner.

more reasonable to suggest that some third influence is clouding the result. One suggestion is that a group of older students has 'plateaued'[34] because of language fossilization (c.f. Selinker and Lamendella, 1979; Vigil and Oller, 1976) at something less than the necessary proficiency to exit this level. Other students who arrived in Canada at the same time as these older students may have already been promoted out of the level at the time the tests were given and thus these younger, more capable students who had been in Canada equally long as the older students would not have been included in the data. In addition, younger, more capable students who arrived in Canada after the older subjects and were included in the data may have surpassed their elders' ability within a single term. Under these conditions it is easy to see that even if there were no relation between the length of time a subject had been in Canada and his language proficiency, for this set of subjects there would be a pattern of correlations similar to that in Table XVI.

Because of the generally complex interrelationship of age and Lot and the language measures, I felt it was necessary to compare solutions including and excluding Lot in a manner similar to that done with age. However, although including Lot in the matrix did simplify the solution (see Appendix M) in

---

[34] At the college where the data was gathered there was, at the time of the research, a class specifically for such older students who did not seem to be progessing in the regular classes. Because of budget constraints this was offered at only one time during the day. It is reasonable to suppose that similar students were attending the regular beginners classes at the other three times during the day.

terms of statistical complexity, it did not improve the interpretability of the factors and so Lot was not included in the matrix in the investigation of the divisibility hypotheses.

5.2.3 First Language

Table XVII presents the results of the investigation of the association of first language with test scores. These results strongly support the argument for the necessity of analyzing the Chinese and non-Chinese separately in the factor analyses. The t-values and associated tests of significance are given in Table XVII, not as a means of accepting or rejecting hypotheses but, as Kruskell (1968, p. 238) expresses it, "as a means of measuring the surprisingness of the observed..." patterns in the data. The results are somewhat surprising. First, in general, the data suggest that something other than construction-induced bias is influencing the variables. The three tests (Listcomp, Liststru, Readvoc) that motivated a contrast between Chinese and non-Chinese speakers are marked with an asterisk.[35] Two of the three particular tests, Readvoc and Listvoc, do show statisically significant differences in means. However, the third, Liststru, which had been subjected to the most extensive revision, does not. Furthermore, Comp, which is not multiple choice and therefore not susceptible to

[35] In these tests, the selection of items and distractors had been based on item-response statistics gathered on samples from the same general population as the research itself. Liststru in particular was composed of items that had undergone several revisions.

Table XVII - Analysis of means grouped by language

VARIABLE		N OF CASES	MEAN	S.D.	t-VALUE	DGRS FRDM	2TAIL PROB.
*READVOC	CHINESE	100	15.21	4.16	3.82	144	0.000
	OTHERS	46	17.86	3.29			
COMP	CHINESE	116	14.16	4.34	2.77	173	0.006
	OTHERS	59	16.24	5.31			
*LISTVOC	CHINESE	110	7.50	2.95	3.21	165	0.002
	OTHERS	57	9.10	3.26			
ORAL	CHINESE	122	14.05	2.46	2.29	179	0.023
	OTHERS	59	15.00	2.87			
LISTCOMP	CHINESE	122	12.43	3.44	2.28	179	0.024
	OTHERS	59	13.67	3.42			
CONCOM	CHINESE	122	7.09	1.87	2.00	179	0.047
	OTHERS	59	7.68	1.89			
LISTBOW	CHINESE	108	14.93	4.49	1.34	155	0.182
	OTHERS	49	16.00	4.86			
ERRCOR20	CHINESE	101	7.38	3.61	1.49	151	0.139
	OTHERS	52	8.32	3.88			
*LISTSTRU	CHINESE	106	12.10	3.51	1.12	153	0.264
	OTHERS	49	12.81	4.02			
READSTRU	CHINESE	111	18.70	3.87	1.11	162	0.269
	OTHERS	53	19.43	4.10			
LISTPHON	CHINESE	92	45.89	9.97	0.80	132	0.424
	OTHERS	42	47.28	7.69			
ERRCOR10	CHINESE	122	5.87	2.46	0.41	179	0.679
	OTHERS	59	6.03	2.22			
LOT	CHINESE	110	31.36	32.99	0.33	155	0.738
	OTHERS	47	29.46	31.32			
AGE	CHINESE	116	31.68	11.80	0.23	166	0.819
	OTHERS	52	32.15	13.46			

*- Multiple choice tests which underwent extensive revision.

construction-induced bias, also displays a statistically significant difference in means. Thus, the pattern is not that predicted by the hypothesized construction-induced bias.

The second 'surprising' fact about Table XVII is that the mean of the Chinese speakers is lower on all of the tests. That these two linguistically defined groups appear to differ in proficiency does not necessarily indicate that the difference results from the difference in language. It could indicate that the two groups differed significantly on some other demographic variable such as level of education. Table XVII does suggest though, that the source of the difference is not linked to age or Lot. The means of the two groups on these two variables are very close in value. Resolution of the exact nature of the source of the differences will need further research and some suggestions regarding this will be made in Chapter Six.

5.2.4 Sex

Table XVIII presents the results of the investigation of the effect of sex on test scores. It presents no evidence to suggest that in this research sex would be linked to any factors that might arise in the factor analysis. The means and standard deviations suggest homogeneity of the two samples. Although two of the differences in means (Listphon, Lot) do approach statistical significance, in a comparison of this many means it is more appropriate to consider these as random events than to attempt to interpret them. In consequence, the analysis of the variable sex is not pursued further.

Table XVIII - Analysis of means, subjects grouped by sex

VARIABLE		N OF CASES	MEAN	S.D.	t-VALUE	DGRS FRDM	2-TAIL PROB.
COMP	F	75	15.26	4.95	0.80	162	0.427
	M	89	14.66	4.59			
LEV1F1	F	69	18.89	3.90	-0.35	150	0.728
	M	83	19.12	3.90			
ERRCOR10	F	76	6.01	2.39	0.35	167	0.723
	M	93	5.88	2.39			
ERRCOR20	F	67	7.79	3.73	0.38	142	0.701
	M	77	7.55	3.51			
CONCOM	F	76	7.42	1.64	0.85	167	0.397
	M	93	7.18	2.02			
ORAL	F	76	14.57	2.60	0.94	167	0.351
	M	93	14.19	2.64			
READVOC	F	59	15.72	4.27	-1.01	132	0.317
	M	75	16.42	3.74			
LISTCOMP	F	76	12.97	3.26	0.29	167	0.769
	M	93	12.81	3.57			
LISTVOC	F	71	8.01	3.22	-0.36	155	0.717
	M	86	8.19	3.08			
LISTPHON	F	60	44.51	10.78	-1.95	123	0.054
	M	65	47.81	8.04			
LISTSTRU	F	66	12.53	3.36	0.21	141	0.835
	M	77	12.40	3.87			
LISTBOW	F	65	15.44	4.21	0.16	145	0.876
	M	82	15.32	4.72			
LOT	F	69	38.04	42.60	2.49	154	0.014
	M	87	25.18	19.87			
AGE	F	73	32.35	10.93	0.42	161	0.675
	M	90	31.53	13.54			

VI. SUMMARY, CONCLUSIONS, AND IMPLICATIONS

6.1 Summary

This study investigated the interrelationships among twelve English language measures and four demographic variables in an attempt to answer the questions 'Is second language proficiency divisible into components and if so what are the components?' The data on the variables were gathered on adult ESL learners in a language course at a community college. The language measures came from three sources. Six were constructed specifically for the research (Liststru, Readvoc, Listvoc, Errcorr20, Listbow, Listphon); four were subtests used in a progress assessment battery at the college (Concom, Listcomp, Errcorr10, Oral); and two were measures used in conjunction with the assessment tests but developed independently as separate projects (Comp, Readstru).

In order to distinguish between linguistic and possible non-linguistic sources of variation in the students' scores, information on four demographic variables (age, sex, length of time in the country, and first language) were also gathered.

During the course of the analysis, it became clear that the effects of age and first language were influencing the relationships among the variables. In addition, there was a high, positive correlation between age and the length of time the subjects had been in Canada. As a result of these findings, the design of the analysis was extended to account for and clarify the effects these variables had on the language

measures. This was done by including age in the correlation matrix used for analysis and by analyzing two subsets of the data, Chinese speakers and non-Chinese speakers, independently of each other. larger group for further analysis.

The principal method of analysis in the investigation of the divisibility hypotheses was that recommended by Hakstian and Bay(1973): image analysis followed by an oblique transformation (Harris-Kaiser independent clusters). The interpretation of the factors focussed on the mode (in particular listening and reading) and content (in particular vocabulary and grammar) of the tests and on the theoretical effects of the demographic variable age. The statistical methods used in the four subsidiary problems were comparison of group means (sex and first language) and correlations (age and Lot) with the language measures.

6.1.1 The Factor Analyses And Interpretation

In the analysis related to the divisibility hypotheses, the data were treated first as a complete (181-subjects) set and then divided into two groups, Chinese speakers (121 subjects) and non-Chinese speakers (60 subjects). In each of the three analyses, the solution indicated three underlying factors influencing the language variables: a structure or grammar factor, a vocabulary or word knowledge factor and an age-related factor. The specific interpretation of the age-related factor is different according to the solution: for the two larger sets it appears to be related to hearing, in the smallest set (non-Chinese speakers) it is more complicated and can be interpreted

as either a bi-polar listening/hearing factor or as a 'speed of processing' factor. Although the distinct nature of the three factors is apparent in each solution, the high correlation between the factors prevents their unqualified interpretation as traits which operate independently.

The clearest and most interpretable factor to appear in each of the solutions is a 'structure' or grammar factor. In each solution, three of the structure-content tests (Readstru, Errcorr10, and Errcorr20) clustered together. The strongest identification of this structure factor is in the solution for the Chinese-speaking subset. In this analysis, all four of the measures designated as structure tests (the three mentioned previously and Liststru) have high coefficients on Factor I and negligible coefficients on Factors II and III. Furthermore, those measures (Concom and Comp) that also loaded on the grammar factor are consistent with the interpretation since the evaluation method in both of these tests includes consideration of grammar.

In addition to the structure factor, a clustering of variables that is interpretable as a vocabulary factor appears in each of the three solutions. This cluster is identifiable by the presence of moderate-to-high coefficients from the two vocabulary tests, Listvoc and Readvoc. However, what strengthens this interpretation is that the two tests are in different modes and therefore the commonality cannot be attributed to modality. Further strength is given to the vocabulary interpretation by the contrasting modes yet

comparable content of other tests that have significant coefficients on this factor. For example, Comp, Oral, and Listcomp are all quite different methods of testing, yet clearly performance on all three will be positively influenced by recognition of or correct use of words.

The third factor in each set was identifiable by the high coefficient of age, though the interpretation of the factor varies depending on data set. In the complete set and in the Chinese-speaking set, age clustered with Listphon and Listbow. Certain features of these tests put a particulary heavy load on the students' hearing ability. First, Listphon is a listening test in which the student must distinguish between minimal pairs [36] which are presented devoid of context. Second, the items in Listbow are spoken at near-natural speed, unlike the other listening tests. Finally, while the items in the other tests are repeated, in these two tests they are not. These aspects of the tests support the suggestion that the age-related factor in the solution for the combined groups and for the Chinese-speakers is best interpreted as a hearing factor.

The age-related factor in the solution for the non-Chinese speakers was different from the solutions for the other two data sets in that age clustered with four listening tests (Liststru, Listcomp, Listphon, Listbow), the oral test and to some extent the composition test. One explanation of this clustering is

---

[36] A minimal pair is a pair of words which differ in only one phoneme, e.g., pin and pen.

that it represents a chance[37] constellation of tests and does not indicate a true relationship between any particular pair of variables. However, since four out of five of the listening tests have their highest coefficients on this factor, if it is not an artifact of chance, it is clearly related to listening mode. Because of this and because the coefficients of all of these tests are opposite in sign to that of age, it is reasonable to interpret it as a bi-polar listening/hearing factor.

Another possible interpretation of the age-related factor takes into account the significant coefficients of Comp (.47) and Oral (.62) on this factor. This interpretation postulates a 'speed of processing' dimension. According to this explanation, in tests which are constrained by time limits, the greatest source of variation among students is their differing ability to integrate all of their linguistic components of 'proficiency' quickly. Listening tests and oral interviews (in which the speed of the language input or stimulus is not controlled by the subject) and time-limited, in-class compositions would exhibit the common influence of such a factor.

While the interpretation of the different clusters in the different solutions is straight forward, their strength is not such that these clusters can be said to represent strong, independent aspects of language proficiency or that the language skill of these sets of subjects is characterized by distinct sub-skills that account for most of the variance in the tests.

[37] As mentioned earlier, the solution for this group is subject to criticism on statistical grounds because of the size of the sample.

In each of the solutions, all three of the factors which were derived were highly correlated.[38] These correlations have two possible explanations. First, they may indicate that in fact only a single dominant factor is influencing the language variables despite the agreement of the separate criteria used to resolve the problem of the number of factors. Three factors were clearly indicated both by the Harris-Kaiser criteria of the number of eigenvalues greater than one in a principal components solution and by the method of inspecting a varimax rotation of a full image analysis for the number of factors with significant loadings. A second possible explanation is that the tests themselves overlap in tapping not only a large general factor but also various combinations of other underlying factors to such a degree that the variables are too complex to produce any solution that is both simple in structure and yet still uncorrelated.

6.1.2 The Demographic Variables

The subsidiary analyses of the four demographic variables suggest that the age and first language of the subjects (as represented by the dichotomy Chinese speakers and non-Chinese speakers) are strongly associated with performance on the language tests. The sex of the subjects, on the other hand,

---

[38] In terms of individuals, the correlation of factors suggests that those who performed well in the tests of one cluster also tended to do well in the tests of another cluster. In the combined groups, the age-related factor was negatively correlated with the two linguistic factors. This indicates that subjects who were older or did poorly on the tests in this cluster tended to do poorly on most tests.

does not appear to be associated with language performance at all. The variable Lot (length of time in Canada), possibly because of its correlation of .42 with age, showed a complicated and ambiguous set of relationships with the language variables.

6.1.3 Age

The variable age appears to be an indirect measure of one or more underlying factors which inhibit performance on the language tests and possibly language acquisition itself. As mentioned earlier, one of the inhibiting factors may be hearing while another could be a slowing down of mental processes in general and language processes (speed of processing) in particular.

6.1.4 First Language

When subjects were categorized as Chinese-speakers and non-Chinese speakers and the data re-analyzed, the means of the Chinese speaking group were lower on all twelve language variables. This finding may not be generalizable to the larger population of language learners because the category 'Chinese-speakers' may be biased by a covert factor, previous education for example. However, within this set of data, the differences between the two groups was of sufficient size to warrant separate analyses of the two groups.

6.1.5 Length Of Time In Canada

The correlation between the measure of the length of time the students had been in Canada and the language scores indicated more of a link to age than I had expected. I have suggested that this statistical link is an artifact of the language program at the college rather than a general demographic connection. As a result of the testing and promotion system at the college, there is a tendency for less capable students to be moved along to and then stopped and held at the proficiency level in the program at which this research was done. Many of these slower students are older and consequently there is probably in the sample an unrepresentatively high proportion of older students who have been in Canada a longer time time than the younger students have. This study suggests potential differences, but relationship between the length of time in Canada and language acquistion will need to be re-addressed in other research.

6.2 Conclusions

Of Oller's (1979a) three divisibility hypotheses, this study tends to support H3 or the model of a general factor plus small specific factors. First, the high correlations of the factors in each solution suggest that a strong global or general proficiency factor is operating, causing moderate correlations among all of the language measures. Second, the consistent emergence of the grammar, the vocabulary, and the age-related factors in each solution argues very strongly for the concept of

multiple, identifiable, specific factors underlying the data, factors which must be taken into account when developing models of either language proficiency or performance on language tests. A proficiency model based on the analyses of the combined group or of the Chinese-speakers subset would include a general factor, a grammar factor and a vocabulary factor. In these two analyses, I believe, the age-related factors reflect an underlying physiological rather than linguistic factor and consquently should not be included in a model of language proficiency. If it were desired to explain language test performance then, certainly, the age-related factor would need to be more clearly defined and then added to the proficiency model.

A model to fit the non-Chinese speaker subset cannot, of course, be fully defined until some better identification of the age-related factor in this set is available. However, this model, too, would include a large general factor and the two specific factors, vocabulary and structure. The age-related factor in the solution for the subset of non-Chinese speakers suggested two interpretations: a true listening-skill factor or a speed of processing factor. If either or both of these represent real factors, then a language proficiency model must include them, too.

This research extends Powers (1982) contention that it is necessary to describe the sample clearly in a divisibility study. Not only does it appear necessary to describe the sample, but it seems crucial to at least explore the

relationships between the language variables and the particular demographic and experiential parameters of a sample. In this study, the age and first language of the subjects and the length of time they had been in Canada were clearly linked to some or all of the language variables. The addition of age actually helped clarify the clusters of linguistic variables. Furthermore, since the solutions for the two subsets were not identical, it can be inferred that the significance of the many underlying factors related to language proficiency may vary according to changes in the nature of the sample. It is clearly in the interest of language acquistion and language testing research to know which factors are stable through the whole population and which factors lose or gain signicance depending on characteristics of the sample.

6.3 Implications And Suggestions For Future Research

The findings of this study have clear implications for future research design. First, they show necessity of fundamentally pure tests if orthogonal factors are to be derived. Almost as important, they indicate that non-linguistic variables must be taken into account if a clear linguistic solution is desired. The specific non-linguistic variables suggested by the research are age, hearing, and any categorical variables such as first language or level of education that may divide the sample into groups that perform significantly differently from each other.

6.3.1 The Correlation Of The Factors

One reason that I have suggested for the correlation of the factors in the factor analyses is the complexity of the tests. In order to establish a dichotomy between grammar and vocabulary, a variety of tests will need to be constucted that put as great a load as possible on one trait while remaining as free as possible of the influence of the other. For vocabulary this might include a simple synonym/antonymn test, or even a test in which students list as many words as they know related to some subject (e.g., parts of the body, kitchen utensils etc.). Still another structure-free vocabulary test would be of the category/example type where students indicate the word that 'doesn't fit.' Creating a vocabulary-free grammar test seems impossible but, by ensuring that all vocabulary used in the grammar tests consists of simple, high frequency words, part of this problem can be solved. The structure tests used in the current research follow this approach to some extent.

Still another way to measure both grammar and vocabulary independently may be to evaluate a composition in two objective ways: first, by taking some simple measure of grammatical accuracy and then by using a measure of diversity of vocabulary. The measure of grammatical accuracy might merely be based on the

number or percentage of correct sentences.[39] For a vocabulary measure, a word frequency count may suffice. Such a count is performed quickly by a computer. There are already numerous word processing programs that will not only count and list the number of different words to appear in a passage but also check spelling.

6.3.2 Age

Certainly in studies where the age range is as broad as it is in this, some account must be taken of the effect of age on the different variables in the study. In this study, not only did all language measures correlate negatively with age, certain of them were affected more than others. By ignoring age and omitting it from the equation, a researcher runs a strong risk of keeping a non-linguistic factor in a matrix but providing no way of identifying it in a solution.

6.3.3 Hearing

In this research I have tentatively linked the age-related factor to hearing. The evidence to support this is not conclusive, but it seems to warrant further studies into this particular aspect of language learning. Clearly, a hearing measure is needed in research where listening tests are

---

[39] At the level of language competence of the subjects of this research, it might be appropriate to ignore punctuation and spelling and to include partial marks for correct clauses. By being too strict there may not be a wide enough spread in scores for the measure to be useful. That is to say, while two compositions may contain an equal percentage of correct sentences, one may have far more 'almost' correct sentences than the other.

involved, particularly where the upper bound of age is as high as it is in this study. Comparing hearing capability with performance on any test (and on progress in language acquisition in general) may prove instructive as well, for obviously being hard of hearing would affect aural learning in general, not only performance on a listening test. Such research would not only be mandatory in the development of theories concerning listening but may also be very useful in counselling adult learners.

6.3.4 First Language

Identification of large, homogeneous subgroups can clarify a study. In this study the homogenous group that was identified was 'Chinese-speakers.' There may, however be more effective ways of grouping subjects or of defining variables that will clarify the apparent differences between groups. For example, although treating the subset of Chinese speakers separately gave a clearer factor solution than treating the entire group did, it may have been even more effective to have obtained an approximate measure of the subjects' exposure to formal education. A number of my Chinese students have commented on the fact that the Cultural Revolution disrupted their education. If a lack of formal education impedes language acquisition, then possibly some subgroup of Chinese accounts for the poorer performance of the group as a whole.

6.3.5 Length Of Time In An English Speaking Environment

The hypothesis that the length of time a subject spends in an English speaking environment is related more strongly to some components of English proficiency than to others is too strong intuitively to abandon. Powers' (1982) interpretation of his earlier study (Swinton and Powers, 1980) gives indirect support to this concept. He has suggested that a vocabulary factor is most influenced by experience and exposure. Although the research was done on subjects who had studied English as a foreign language (EFL), it is reasonable to expect a similar result for ESL subjects. The reason my reseach did not clarify the issues was that the relationship between the linguistic, variables and Lot had been confounded by Lot's relation with age. This appears to have been a result of a group of older students who had become 'stuck' at the one level. If possible, a sample should be taken from a program in which students remain only a limited amount of time.[40] The suggestion for further research is to deal with a program that does not have proficiency barriers at various points in the program.

---

[40] One such program would be the five month, Canada Manpower sponsored language classes given in various colleges and other centres across the country.

BIBLIOGRAPHY

1. Allen, J.P.B., and Alan Davies (eds.) 1979. Testing and Experimental Methods. London:Oxford University Press

2. Bachman, L.F., and A.S. Palmer. 1982. "The Construct Validity of Some Components of Communicative Proficiency." TESOL Quarterly 16:449-465

3. Bachman, L.F., and A.S. Palmer. 1981. "The Construct Validation of the FSI Oral Interview." Language Learning 31:67-86

4. Bowen, J.D. 1975. An Experimental Integrative test of English Work Papers in Teaching English as a Second Language, Vol. 9 Los Angeles: University of California, Dept. of English

5. Borg, W.R., and M.D. Gall. 1979. Educational Research: An Introduction New York: Longman, Inc.

6. Canale, M., and M. Swain. 1980. "Theoretical Bases of Communicative Approaches to Second language Teaching and Testing." Applied Linguistics 1:1-47

7. Carroll, B.J. 1980. Testing Communicative Performance Oxford: Pergamon Press

8. Carroll, J.B. 1968. "The Psychology of Language Testing." In Alan Davies, ed. Language Testing Symposium: A Psycholinguistic Approach. London:Oxford University Press, 1968,p. 46-69

9. Cattell, R.B. 1962. "The Basis of Recognition and Interpretation of Factors." Educational and Psychological Measurement 22:667-697

10. Cattell, R.B., 1958. "Extracting the Correct Number of Factors in Factor Analysis" Educational and Psychological Measurement 18:791-837

11. Cooper, Robert L. 1968. "An Elaborated Language Testing Model." Language Learning 3: 57-65

12. Davies, Alan (ed.) 1968. Language Testing Symposium: A Psycholinguistic Approach. London:Oxford University Press

13. Diederich, P.B. 1974. Measuring Growth in English. Champaign, Illinois: National Council of Teachers of English

14. Ebel, Robert L. 1972. Essentials of Educational Measurement Englewood Cliffs, New Jersey: Prentice-Hall

Inc.

15. Farhady, H. 1979. "The Disjunctive Fallacy Between Discrete Point and Integrative Tests." TESOL Quarterly 13:347-357

16. Flahive, Douglas E. 1980. "Separating the g Factor from Reading Comprehension." In John W. Oller Jr., and Kyle Perkins, eds. Research in Language Testing Massachusetts: Newbury House, 1980, p. 34-46

17. Gardner, R.C., and L. Gliksman. 1982. " On 'Gardner on Affect': A Discussion of Validity as it Relates to the Attitude/Motivation Test Battery: A Response From Gardner." Language Learning 32:191-200

18. Gorsuch, R.L. 1974. Factor Analysis Philadelphia: Saunders

19. Hakstian, A. Ralph, and Kyung S. Bay. 1973. User's Manual to Accompany the Alberta General Factor Analysis Program (AGFAP). Alberta:University of Alberta.

20. Hakstian, A.R., and V.J. Muller. 1973. "Some Notes on the Number of Factors Problem." Multivariate Behavioral Research 4:461-475

21. Harman, H.H. 1976. Modern Factor Analysis (Third Edition) Chicago: The University of Chicago Press

22. Harris, David P. 1968. Testing English as a Second Language. New York:McGraw-Hill

23. Hendricks, D. George Scholz, Randon Sperling, Marianne Johnson and Lela Vandenberg. 1980. "Oral Proficiency Testing in an Intensive English Language Program." In John W. Oller Jr., and Kyle Perkins, eds. Research in Language Testing Massachusetts: Newbury House, 1980, p. 77-90

24. Ingram, E. 1978. "The Psycholinguistic Basis." in Bernard Spolsky, ed. Papers in Applied Linguistics: Advances in Language Testing Series:2 Approaches to Language Testing, Arlington Verginia: Center for Applied Linguistics, 1978, p. 39-58

25. Johansson, S. 1973. "Partial Dictation as a Test of Foreign Language proficiency." Swedish-English Contrastive Studies Report No. 3, Department of English, Lund University, Sweden

26. Joreskog, K.G. 1969. " A General Approach to Confirmatory Maximum Likelihood Factor Analysis." Psychometrica 32:443-482

27. Joreskog, K.G. 1978. "Structural Analysis of Covariance and Correlation Matrices" Psychometrica 43:443-477

28. Kendall, M. 1980. Multivariate Analysis. London : Charles Griffin and Company

29. Krishnaiah, P.R. and J.C. Lee. 1980. "Likelihood Ratio Tests for Mean Vectors and Covariance Matrices" in P.R. Krishnaiah, Handbook of Statistics, Vol.1 Analysis of Variance. Amsterdam, New York: North-Holland Publishing Company, 1980

30. Kruskell, J. 1968. "Tests of Significance" in David L. Sills, ed. Encyclopedia of the Social Sciences Volume 14, The MacMillan Company and The Free Press, 1968, p. 238-249

31. Magnusson, D. 1967. Test Theory. Addison Wesley

32. Mullins, K.A. 1980. "Rater Reliability and Oral Proficiency Evaluations." In John W. Oller Jr., and Kyle Perkins, eds. Research in Language Testing Massachusetts: Newbury House, 1980, p. 91-101

33. Munby, J.L. 1978. Communicative Syllabus Design. Cambridge University Press

34. Nilsen, D.F., and A.P. Nilsen. 1973. Pronunciation Contrasts in English New York: Regents Publishing Company.

35. Nie, N., C.H. Hull, J.G. Jenkins, K. Steinbrenner, D.H. Bent. 1975. SPSS: Statistical Package for the Social Sciences. McGraw-Hill Book Company.

36. Nunnally, J.C. 1978. Psychometric Theory. McGraw-Hill Book Company

37. Oller, John W. Jr. 1976a. " Language Testing Today: an Interview with John Oller." English Teaching Forum. July. P. 22-27

38. Oller, John W. Jr. 1976b." A Program for Language Testing Research" Language Learning. 4: 141-165

39. Oller, John W. Jr., 1978. "Pragmatics and Language Testing" in Bernard Spolsky, ed. Papers in Applied Linguistics: Advances in Language Testing Series:2 Approaches to Language Testing, Arlington Virginia: Center for Applied Linguistics, 1978, p. 39-58

40. Oller, John W. Jr. ed. 1979a. Language Tests at School: A Pragmatic Approach London:Longman

41. Oller, John W. Jr. 1979b. (Class notes from course

Advanced Issues in Language Testing at First Annual TESOL Summer Institute)

42. Oller, John W. Jr. 1981. "Language Testing Research (1979-1980)" in R.B. Kaplan, General ed., Randall L. Jones, and G.R. Tucker Co-editors. Annual Review of Applied Linguistics 1980. Rowley, Massachusetts : Newbury House, 1981, p. 124-150

43. Oller, John W. Jr. and Frances Butler Hinofotis. 1980. "Two Mutually Exclusive Hypotheses about Second Language Ability:Indivisible or Partially Divisible Competence." In John W. Oller Jr., and Kyle Perkins, eds. Research in Language Testing Massachusetts: Newbury House, 1980, p. 13-23

44. Oller, John W. Jr., and Kyle Perkins(eds.). 1980. Research in Language Testing Massachusetts:Newbury House 206-229

45. Pike, L.W. 1979. An Evaluation of Alternative Item Formats for Testing English as a Foreign Language TOEFL Research Reports (2) Princeton: Educational Testing Service

46. Powers, D.E. 1982. "Selecting Samples for Testing the Hypothesis of Divisible Versus Unitary Competence in Language Proficiency." Language Learning 32:331-335

47. Scholz, George E., and Celeste M. Scholz. 1979. "Testing in an EFL/ESP Context." in Carlos A. Yorio, Kyle Perkins and Jacquelyn Schacter (eds.) On TESOL '79 The Learner in Focus. Washington, D.C.:Teachers of English to Speakers of Other Languages, 1979, p. 206-209

48. Scholz, George, Debby Hendricks, Randon Spurling, Marianne Johnson, and Lela Vandenberg. 1980. "Is Language Ability Divisible or Unitary? A Factor Analysis of 22 English Language Proficiency Tests." in John W. Oller Jr., and Kyle Perkins, eds. Research in Language Testing Massachusetts: Newbury House, 1980, p. 24-33

49. Selinker, L. and J. Lamendella. 1979. "The Role of Extrinsic Feedback in Interlanguage Fossilization.: Language Learning 29:368-375

50. Streiff, Virginia 1978. Relationships among Oral and Written Cloze Scores and Achievement Scores in John W. Oller Jr., and Kyle Perkins (eds.) Language in Education:Testing the Tests. Massachusetts:Newbury House, 1978, p. 65-102

51. Stump, Thomas A. 1976. "Cloze and Dictation Tasks as Predictors of Intelligence and Achievement Scores." in

John W. Oller Jr., and Kyle Perkins (eds.) Language in Education:Testing the Tests. Massachusetts:Newbury House, 1978, p. 65-105

52. 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 6, New Jersey: Educational Testing Service

53. Upshur, John A.1976. Discussion of "A Program for Language Testing Research" Language Learning 4. 167-174

54. Valette, Rebecca M. 1977. Modern Language Testing New York:Harcourt Brace Jovanovich

55. Vigel, J. and J.W. Oller. 1976. "Rule Fossilization: A tentative Model." Language Learning 26:281-295

APPENDIX A - INTRODUCTION, SAMPLE ITEMS, AND SAMPLE ANSWER SHEET FROM LISTENING-STRUCTURE TEST USED IN PILOT STUDY AND MAIN RESEARCH

A. Introduction

Here is another listening exercise. This is a review of Beginners' grammar. Look at your answer sheet and listen to the first example:

My book ***** [41] on the table.

a. are
b. is
c. do

(Repeat)

The answer to that is obviously letter "b"..."is." Did you circle letter 'b?'
Let's try example two:

Bill, where ***** my jacket?

a. you put
b. you did put
c. did you put

(Repeat)

Did you circle letter "c." That is the correct answer.
Now try example three:

He ***** to Eaton's tomorrow, to pick up some shoes.

a. is going
b. will
c. has gone

(Repeat)

The correct answer to that one is "a"..."is going." Did you circle "a?"
Now try example four. It is different.

Yesterday, John ***** to work.

a. going
b. goes

---

[41] In this script, the ***** indicates the sound of a bell.

c. gone

(Repeat)

There is no correct answer. Did you circle no?
If you want, your teacher will play the instructions again.
All right. Let's begin.

B. Example items

(The following are the first five examples from the final version of this test, used in the main study. Each question is repeated.)

1. He couldn't buy a sandwich because he didn't have ***** money.
   a. some
   b. many
   c. enough

2. Hey, don't eat that sandwich. It is *****.
   a. my
   b. I
   c. mine

3. Mrs. Wright can't go with us because ***** car is not working.
   a. She's
   b. hers
   c. her

4. John, ***** you tired last night.
   a. have
   b. were
   c. do

5. Edward, how ****** you come to school last Thursday?
   a. will
   b. do
   c. are

C. Sample answer sheet

The following is a example answer sheet for the first five questions on the listening-structure multiple choice test:

Example 1. a b c NO

Example 2. a b c NO

Example 3. a b c NO

Example 4. a b c NO

1. a b c NO

2. a b c NO

3. a b c NO

4. a b c NO

5. a b c NO

6. a b c NO

7. a b c NO

8. a b c NO

9. a b c NO

APPENDIX B - EXAMPLE ITEMS FROM THE READING VOCABULARY TESTS USED IN THE PILOT STUDY AND THE MAIN RESEARCH

1. He is the thief who _____ all my money.

   a. borrowed
   b. stole
   c. loaned
   d. reduced

2. I was laid off so now I am _____ .

   a. looking for a job
   b. employed
   c. in bed
   d. tired

3. When people buy a house or a car, the first money they pay is the _____ .

   a. monthly payment
   b. principal
   c. interest
   d. down payment

4. My pencil is broken. Could you _____ me yours for a minute?

   a. exchange
   b. offer
   c. lend
   d. borrow

5. Could you tell me your address? I have _____ it.

   a. visited
   b. forgotten
   c. remembered
   d. (no correct answer)

6. This old bicycle is not working. I am going to take it to the bicycle shop for _____ .

   a. a refund
   b. repairs
   c. a mechanic
   d. (no correct answer)

APPENDIX C - INTRODUCTION AND SAMPLE ITEMS FROM LISTENING VOCABULARY TEST USED IN PILOT STUDY AND MAIN RESEARCH

A. Introduction

Hello are you ready for another listening exercise? This is a vocabulary exercise. How many English words do you know? Look at your answer sheet. Now listen to this:

I want to buy a coffee. Would you lend me *****?

a. some water
b. some money
c. your car

Listen again. (Repeat)

The answer to that is obviously letter "b" ... "some money." Did you circle "b?" Let's try example two.

John is in the classroom, reading ****.

a. a movie
b. sandwich
c. a book

(repeat)

Did you circle letter "c." ... "a book?" That's the correct answer. Now try example 3.

I need a **** because I'm going to write a letter.

a. pen
b. shoe
c. doctor

(repeat)

The correct answer to that one is "a"... "pen." Did you circle "a?" Now try example four. It's different.

I'm going to the **** to buy some stamps.

a. bank
b. beach
c. movie

(repeat)

None of the words are correct, are they? Did you circle no for no correct answer? If you want, your teacher will play the examples again.

B. Sample items

The following are the first five items used on the listening vocabulary test in the main research.

1. He wants to save money so he's going to open *****.

   a. a deposit
   b. a check
   c. an account

2. When you borrow money from a bank, you pay *****.

   a. back
   b. cash
   c. interest

3. My son was sick so I made ***** with the doctor.

   a. an appointment
   b. a prescription
   c. a telephone

4. The federal government takes two hundred dollars from my pay every month. I don't like paying *****.

   a. insurance
   b. income
   c. taxes

5. If you want help in a department store, ask the *****.

   a. secretary
   b. teller
   c. deposit

(NOTE: The answer sheet is the same as in the listening structure test.)

APPENDIX D - EXAMPLE ITEMS FROM THE READING GRAMMAR TESTS USED IN THE PILOT STUDY AND THE MAIN RESEARCH

1. He couldn't buy a sandwich because he didn't have _____ money.

   1. some
   2. many
   3. enough
   4. more

2. This is a low priced car. It is _____ the others.

   1. as not expensive as
   2. not as expensive
   3. not as expensive as
   4. as expensive not

3. My mother can't go with us because _____ car is not working.

   1. she
   2. she's
   3. hers
   4. her

4. That car is _____ car of all.

   1. more comfortable
   2. the most comfortable
   3. most comfortable
   4. the more comfortable

5. It is a big class but there aren't _____ in it.

   1. much women
   2. many women
   3. a lot women
   4. some women

6. She _____ come to the meeting tomorrow because she has a dentist appointment.

   1. doesn't
   2. hasn't
   3. won't
   4. couldn't

APPENDIX E - EXAMPLE OF CONVERSATION COMPLETION TYPE OF SUBTEST USED IN ASSESSMENT BATTERIES

COMPLETE THE CONVERSATION

Mary took a suit to the dry cleaners last week. She picked it up this morning. The zipper is broken. She is at the dry cleaners now. She is complaining to the manager.

Manager: Good afternoon. May I help you?

Mary: _____

Manager: What's the matter?

Mary: _____

Manager: Do you have your bill?

Mary: _____

Manager: O.K. We'll repair it for you.

Mary: _____

Manager: It will be ready on Saturday.

Mary: _____

10 marks

APPENDIX G - INTRODUCTION, SAMPLE ITEM, AND SAMPLE ANSWER SHEET FROM LISTENING COMPREHENSION TEST USED IN MAIN RESEARCH

A. INTRODUCTION

Listening test. Listen to these examples and do them with your teacher.
WOMAN: Is this your sweater or John's?
MAN: Let me see, Oh, it's mine.

(Repeat)

Question 1: Who does the sweater belong to?

a. the man
b. John

Question 2: What colour is the sweater?

a. red
b. blue

(Pause 5)

Example 2
WOMAN: That bus is late again.
MAN: Yes it always is when it rains.

(Repeat)

Question 1: What are the man and woman doing?

a. drinking coffee
b. waiting for a bus

Question 2: What is the weather like?

a. cold
b. sunny

(Pause 5)

Instructors, you may play the examples several times. Be sure the students understand

Let's begin.

B. Sample Item from Listening Comprehension test

WOMAN: Would you like some dessert, sir?

MAN: Hmmm, yes, please. What's good today?

WOMAN: Well, there's chocolate cake. We also have cherry pie.

MAN: Cherry pie? Hmmm, no. Give me a piece of the cake.

(Repeat)

Question 1: What does the man want?

a. chocolate cake
b. cherry pie

Question 2: Where are these people?

a. in a restaurant
b. at home

C. Sample Answer Sheet

1. a b no
2. a b no
3. a b no
4. a b no
5. a b no
6. a b no
7. a b no
8. a b no
9. a b no
10. a b no
11. a b no
12. a b no

APPENDIX H - ORAL INTERVIEW GUIDELINES AND SAMPLE SCORING SHEET

Upper Beginners Oral Test -- Score Sheet

Part A	General Comprehension				/5
Part B	'Free' Speaking Fluency			Accuracy	
	0 1 2		+	0 1 2	/4
Part C	Question Making				
	Content/ Production			Accuracy	
	0 1 2		+	0 1 2	/4
Part D	Language Use Appropriateness/ Completeness			Accuracy	
1.	0 1 2		X	0 1/2 1	/2
2.	0 1 2		X	0 1/2 1	/2
3.	0 1 2		X	0 1/2 1	/2
					/20

ORAL TEST Upper Beginners

Part A (one minute)

Use these questions to set the students at ease and to test their general comprehension. Speak in a conversational tone and at regular speed. If the students answer in any way (short, long) that shows comprehension of the questions, give full marks.

1. How are you today?
2. Sit down.
3. What is your name?
4. How do you spell your (first, last) name?
5. Who is you teacher?
6. How long have you been in Upper Beginners?
7. Is this your first interview?
8. Where do you live?
9. (If they give an address, area, ask:) Where is that?

Part B (two minutes)

Ask the student to tell you about ONE of the following:

1. Educational background
2. Employment background
3. Activities on a particular job
4. (For unemployed students who also have little to say about their education...) Day-to-day activities

Some suggested lead-ins.....

1. Did you go to school in .....? Tell me a little about what you studied.
2. How many different jobs have you had? (Have you had several jobs?) Tell me a little about those different jobs.
3. Are you working now? Tell me what you do on your job. (or) Tell me about where you work.
4. (Housewives, young students and some others may not have anything to say about the first three topics, ask ....) What do you do during the day?

Part C

The student must ask at least three relevant questions from the point of view of someone renting an apartment.

You may guide the student as to acceptability of the questions she/he asks and prompt for more.
Acceptable ... Questions relating to rent, number of bedrooms, the floor it's on, nearness to schools/shops, when it is available.

Lead-in ....

You are looking for an apartment. I am the apartment owner. Ask me some important questions about my apartment....

Tester engages in conversation with the student.

Part D (three minutes)

Using the language....Students must make an appropriate response to each of the problems presented. The appropriate response includes: complexity, stress, register, intonation, appropriatness of the utterance to the situation, mood created by the response. Choose one of the three questions for each communicative type.

Apologizing

1. Your friend invited you to go to a movie tomorrow night. You can't go because you have something else to do. What do you say? Tell him/her why.
2. You are sick today and can't go to work. You phone your employer. What do you say to him?
3. You are late for class. Your teacher looks upset. What do you say to her/him? Make an excuse and a promise.

Complaining

1. You bought a hamburger at a take-out restaurant. It is cold and doesn't taste good. What do you say?
2. You took a dress to the dry cleaners last week. When you picked it up it had a button missing and the zipper was broken. What do you say?
3. You bought some milk at the corner store this morning, but it is sour. You take it back to the store. What do you say?

Social Situation

1. Your friend's mother died last week. What do you say to her?

2. Invite me to have a cup of coffee with you after class.

3. Introduce me to your friend.

APPENDIX I - COMPOSITION MARKING GUIDE

Verbal Description for Free-Writing Assessment -- Beginners to College Entry

Level 1: Semantics (Function, Vocabulary, Organization)

Can give (ask for) concrete info. (name, address, phone, place of work, etc.) With a picture series of everyday basic experience for guidance, can write very brief, simple description or report. Can't handle discussion. Vocabulary limited and often inappropriate. Any organization due to picture guidance.

Level 1: Syntax (Interclause)

Can produce some simple sentences (affirmative, negative, interrogative). May attempt simple co-ord (and, or, but, so) and sub-ord (when, because)

(Intraclause)

Demonstrates awareness of past/present/future time but not always correct. Frequent errors of the following types: word order, word-form, fragments, run-ons, pronoun and subject-verb agreement, prepositions and articles.

Level 1: Orthography (Punctuation, Spelling, Readability)

Little or no punctuation or capitalization. Frequent spelling errors in common words. Letters unclear, messy paper. Almost impossible to read.

Level 2: Semantics

Can rearrange stock phrases and patterns to handle basic personal and survival areas. With a verbal rather than pictorial stimulus, some difficulties with describing and reporting in these areas. Can't handle discussion. Vocabulary is high frequency and generally appropriate for these basic areas. Not necessarily well organized.

Level 2: Syntax (Interclause)

Simple sentences generally mastered as well as some success in simple co-ord and sub-ord from Level 1. Other types of sub-ord may be attempted.

Intraclause

Use of past/present/future generally correct. Problems in

use of simple vs. continuous and present perfect. Few problems with word order. Prepositions of time/place/direction generally correct but continuing difficulties with word form, fragments and run-ons, agreement, idiomatic prepositions and articles.

Level 2: Orthography

Some attempt at punctuation. Less than impossible to read. Frequent spelling errors.

Level 3: Semantics

Can handle description and reporting in everyday situations but has some difficulties with discussion. Exhibiting choice about vocabulary which is adequate for informal communication and some use of idioms. Great difficulties with abstract or distant levels of the topic if attempted. Some effort at organization including transitions.

Level 3: Syntax (Interclause)

Simple sentences, co-ord & sub-ord. (adv. + adj.) generally correct. May have difficulties with N. clauses especially from questions. May attempt abridgements and phrases (participle, gerund, infinitive) but unsuccessfully.

(Intraclause)

Past/Pres./Future, contin. and Perfect under control. May have problems with past perf., use of tenses in conditions and sequencing across sentences. Few problems with frags., r.o.'s and agree. Use of articles and common idiomatic prepositions generally correct but difficulties with 'the' deletion, less common id. preps. and word form.

Level 3: Orthography

Punctuation correct most of the time, especially capitalization and period but use of comma may be erratic. Some attempt at paragraphs. Few spelling errors in common words.

Level 4: Semantics

Can handle description, reporting and discussion on everyday level. Beginning to control topic at more abstract or distant levels. Appropriate use of idioms and lower frequency vocab. Some flaws in organization. May contain some redundancy.

Level 4: Syntax (Interclause)

N. clauses, abridgements, phrases generally correct. May attempt absolute constructions, abstract noun phrases and appositive phrases.

(Intraclause)

Few, if any problems with odd use of tenses and sequencing. Few problems with word form. Use of less common id. preps. generally successful. Few, if any problems with articles including 'the' deletion.

Level 4: Orthography

Control of remaining punctuation: commas, colons, semicolons, quotation marks, etc. Spelling correct except for words natives would find difficult. Proper paragraphing.

Level 5: Semantics

Acceptable for college entry. Can handle description, reporting and discussion even at abstract and distant levels. Vocab. and idioms appropriate to the task and vary in frequency... high and low. Clear, logical patterns, adequate development and lack of redundancy. Occasional evidence of unnatural but correct English.

Level 5: Syntax (Interclause)

Can handle a wide variety of grammatical functions and sentence types with few, if any, errors. Absolute constructions, abstract noun phrases and appositive phrases are correct if attempted.

(Intraclause)

Any intra-clause errors probably due to carelessness.

Level 5: Orthography

Beautiful, easy to read. No errors which would produce any misunderstanding or embarrassment.

APPENDIX J - INTRODUCTION, SAMPLE ITEMS, AND SAMPLE ANSWER SHEET FROM PHONEME DISCRIMINATION TEST

A. Introduction

How well can you hear the different sounds of English? In each question you will hear three words. Circle the number of the one that is different.
Listen to example one

meet meet mate

The third word was different. So you should circle three.
Sometimes all of the words are the same.
Look at example two.

meet meet meet

You should circle 's' for same.
Sometimes all of the words are different.
Listen to example three.

meet mate mote

You should circle 'd' for different.
Here are five more easy examples. Do them on your answer sheet.
Example four

steak stock steak

The answer is two. Did you circle two?
Example five:

brick break break

The answer is one. Did you circle one?
Example six:

coin coin coin

They are all the same. Did you circle 's?'
Example seven:

can can coin

Number three is the right answer. Did you circle three?
Example eight:

coin can cane

Those are all different. Did you circle 'd?'

Teachers you may play the examples again. All right. Let's begin.

B. Sample Items

1. gene gin gin
2. swayed swede swayed
3. leaned leaned leaned
4. bit bait bet
5. peck pick peck
6. sling sling sling
7. sting stung sting
8. lace less less
9. pad paid pad
10. lake lake lake

C. Sample answer Sheet

1.	1	2	3	s	d
2.	1	2	3	s	d
3.	1	2	3	s	d
4.	1	2	3	s	d
5.	1	2	3	s	d
6.	1	2	3	s	d
7.	1	2	3	s	d
8.	1	2	3	s	d
9.	1	2	3	s	d
10.	1	2	3	s	d
11.	1	2	3	s	d
12.	1	2	3	s	d
13.	1	2	3	s	d
14.	1	2	3	s	d
15.	1	2	3	s	d

APPENDIX K - INTRODUCTION, SAMPLE ITEMS, AND SAMPLE ANSWER SHEET FROM THE BOWEN-FORMAT LISTENING TEST

A. Introduction

Listening Exercise. When people speak English quickly, some words become shorter. For example, we don't say 'He is going.' we say 'He's going. Sometimes words get pushed together. For example, we don't say 'Is he going?' we say 'Is-he going?' In this exercise, listen carefully to the sentences. Then write only the second word you hear. Look at your answer sheet now, and do the examples with me. The first one has been done for you.

Example A: What's he doing? (Pause 5)
Did you hear "is"? -i-s The sentence is 'What is he doing?'

Example B: Did he leave his book? (Pause)
Did you write 'he' -h-e? The sentence is 'Did he leave his book?'

Example C: What did you do yesterday? (Pause)
The answer is 'did' d-i-d. The sentence is 'What did you do yesterday?"

Example D: What colour is her car? (Pause)
Did you write 'colour?' The sentence is 'What colour is her car?"
Here are four more examples. Do them with your instructor. Write your answers under group two.

E. Where did you go last week?

F. Did he do his homework?

G. When is he coming?

H. What kind of ice-cream do you like?

B Sample items from test

1. She's leaving tomorrow morning.
2. Did he pay for the dinner?
3. John and Nancy are coming to the party tonight.
4. He's tired of that, isn't he?
5. Is your brother coming to pick you up?
6. What do you think the weather will be like tomorrow?
7. What did he do all day at the library?
8. Well, there are no more books here.

C. Sample answer sheet

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____

APPENDIX L - EXPERIMENTAL LISTENING TEST USED IN PRELIMINARY STUDY

Introduction

Instructors, be sure the students are looking at the columns of NGs and OKs.

In this test you will hear some short conversations between a woman and a man. You will hear them only once. After each conversation decide if the man's answer is OK or not OK. If it is OK, circle OK on the answer sheet. If it is not OK, circle NG for no good.

Here are four examples.

Example one.

Good morning. How are you?
Fine and you?
(4 second pause)

The man's answer is OK. Did you circle OK?

Example two.

What's the weather like today?
It's Monday
(4 second pause)

The man's answer is certainly not OK. Did you circle NG for no good?

Example three.

Where is my book?
On the table.
(4 second pause)

The man's answer is short but it is good. Did you circle OK?

Example four.

Do you go to school every day?
Yes I do, but only on Tuesdays and Thursdays.
(4 second pause)

The first part of the man's answer is OK but the second part is no good. Did you circle NG? Teachers, be sure the students understand. You may play the examples again.

B. Items

1. When is the party going to be?
   Next Tuesday.

2. Why didn't you come yesterday?
   I had to see a doctor.

3. How long is your holiday?
   Two or three kilometers.

4. Is it warm enough in here?
   Yes it is. I need my coat.

5. Is this the first time you've been to Vancouver?
   Yes, that's right. I have only been here once before.

6. I'm not feeling very well.
   Oh really, what happened to you?

7. Why are your hands dirty?
   Because I've just finished cleaning the cupboards.

8. I have never played cards.
   It is like tennis.

9. This sweater isn't big enough for me.
   Yes, it is too big, isn't it?

10. I want five seventeen cent stamps.
    I'm sorry. We have just run out.

Sample answer sheet

Example 1. OK NG

Example 2. OK NG

Example 3. OK NG

Example 4. OK NG

1. OK NG

2. OK NG

3. OK NG

4. OK NG

5. OK NG

6. OK NG

7. OK NG

8. OK NG

9. OK NG

10. OK NG

APPENDIX M - THE AUXILIARY FACTOR ANALYSES

The purpose of this series of analyses was to investigate the effect on the factor matrix of the two variables of undetermined reliability (Concom and Oral) and of the two demographic variables age and Lot. and to determine which combination of language and demographic variables to use in arriving at preferred solutions. Missing data was replaced with group means and the method of factor analysis was principal components followed by a varimax rotation. The results of this series of factor analyses --done on six subsets of the variables-- are presented in Table XIX.

In the series of six analyses, each of which included eleven or more variables, all solutions produced three factors. The most important fact about Concom and Oral was that they did not create or define factors when they were introduced. That is to say, the pattern of loadings was substantially the same with or without either of these two variables. The conversation completion (Concom) consistently clustered with the three structure tests and did not reveal any complexity. The loadings of Oral on the other hand did change depending on the presence or absence of age. This is not surprising given the correlation (-0.36) of age and Oral. I did not consider this vacillation as inherent weakness or unreliablity on the part of Oral because Comp, Errcorr20, Listcomp, and Listbow were also affected by the addition/deletion of this variable. In general, the inclusion of the two demographic variables simplified the solutions. That is, when these variables were included, the number of variables loading on three factors decreased. In the solution for the twelve linguistic measures there were two variables, Comp and Listcomp which "spread out" over all three factors. When age or Lot was included, these reduced to two-variable complexity, and it became clear that there were in fact only two linguistic factors influencing these two and the rest of the language variables. The third major factor was a demographically defined one. By adding age to the matrix of language variables, Thurstone's criteria of simple structure was more clearly met. However, when both age and Lot were included, the solution lost some of its interpretive power in that the coefficients for Readvoc and Liststru became more evenly distributed on two factors. Because of this, the final solutions did not incorporate Lot.

Table XIX - Factor by factor comparison of subsets of the variables (principal components followed by varimax rotation)

			FACTOR I			
COMP	.62	.64	.57	.53	.67	.64
READSTRU	.78	.69	.70	.72	.71	.70
ERRCOR10	.79	.76	.80	.77	.77	.75
ERRCOR20	.76	.72	.65	.63	.74	.72
CONCOM	.63	**	**	.47	**	**
ORAL	.18	.26	.24	.22	**	**
READVOC	.38	.37	.29	.31	.38	.35
LISTCOMP	.38	.39	.35	.36	.43	.40
LISTVOC	.05	.14	.11	.08	.14	.08
LISTPHON	.25	.27	.20	.21	.27	.26
LISTSTRU	.42	.42	.39	.35	.45	.43
LISTBOW	.25	.33	.26	.16	.37	.35
LOT	-.09	-.05	-.02	**	-.03	**
AGE	.08	.04	**	**	.00	.01
			FACTOR II			
COMP	.06	.07	.09	.41	.09	.16
READSTRU	.16	.17	.14	.27	.16	.19
ERRCOR10	.11	.12	.07	.22	.12	.17
ERRCOR20	.05	.06	.12	.31	.05	.06
CONCOM	.08	**	**	.12	**	**
ORAL	.30	.27	.16	.47	**	**
READVOC	-.06	-.02	.02	.18	.01	.02
LISTCOMP	.14	.15	.06	.38	.18	.28
LISTVOC	-.17	-.12	-.21	-.10	.07	.06
LISTPHON	.66	.51	.58	.39	.53	.44
LISTSTRU	.13	.13	.10	.35	.14	.21
LISTBOW	.44	.39	.34	.76	.40	.47
LOT	-.75	-.58	-.60	**	-.58	**
AGE	-.79	-.74	**	**	-.72	-.82
			FACTOR III			
COMP	.51	.44	.52	.38	.40	.42
READSTRU	.16	.17	.21	.10	.10	.11
ERRCOR10	.23	.20	.23	.19	.19	.17
ERRCOR20	.24	.21	.30	.19	.15	.21
CONCOM	.05	**	**	.15	**	**
ORAL	.63	.50	.52	.31	**	**
READVOC	.49	.40	.47	.45	.40	.48
LISTCOMP	.66	.60	.63	.49	.54	.52
LISTVOC	.78	.66	.65	.78	.74	.75
LISTPHON	.12	.09	.21	.02	.07	.05
LISTSTRU	.39	.31	.36	.22	.27	.26
LISTBOW	.55	.44	.51	.19	.36	.29
LOT	.28	.20	.11	**	.21	**
AGE	-.28	-.25	**	**	-.17	-.03

APPENDIX N - CORRELATION MATRIX OF ALL VARIABLES

	COMP	READST	ERCO10	ERCO20	CONCOM	ORAL
COMP	1.00					
READSTRU	0.543	1.00				
ERRCOR10	0.588	0.678	1.00			
ERRCOR20	0.610	0.606	0.639	1.00		
CONCOM	0.328	0.446	0.385	0.400	1.00	
ORAL	0.421	0.360	0.294	0.382	0.209	1.00
READVOC	0.481	0.360	0.393	0.429	0.285	0.336
LISTCOMP	0.511	0.415	0.433	0.450	0.342	0.454
LISTVOC	0.401	0.163	0.259	0.228	0.153	0.301
LISTPHON	0.365	0.273	0.337	0.286	0.153	0.252
LISTSTRU	0.457	0.385	0.424	0.471	0.199	0.305
LISTBOW	0.507	0.404	0.345	0.427	0.220	0.475
LOT	0.025	-0.091	-0.038	-0.070	-0.103	-0.025
AGE	-0.110	-0.193	-0.176	-0.036	-0.109	-0.367

	READVO	LISTCO	LISTVO	LISTPH
READVOC	1.00			
LISTCOMP	0.421	1.00		
LISTVOC	0.476	0.454	1.00	
LISTPHON	0.226	0.241	0.075	1.00
LISTSTRU	0.256	0.457	0.278	0.236
LISTBOW	0.327	0.442	0.285	0.418
LOT	0.013	0.003	0.213	-0.438
AGE	0.004	-0.276	-0.106	-0.413

	LISTSR	LISTB	LOT	AGE
LISTSTRU	1.00			
LISTBOW	0.407	1.00		
LOT	-0.021	-0.135	1.00	
AGE	-0.171	-0.434	0.416	1.00