2
2001

This is to certify that the

dissertation entitled

INVESTIGATING GROWTH TRAJECTORIES
ON ENGLISH AS A SECOND LANGUAGE LISTENING
AND READING COMPREHENSION TESTS

presented by

H. Gary Cook

has been accepted towards fulfillment
of the requirements for

PH.D. degree in Education

William Mehrens
Major professor

Date April 12, 2001

DATE DUE	DATE DUE	DATE DUE
JUL 1 6 2005 C J 1 8 0		
FEB 1 9 2006 C 9 1 8 0		
NOV 2 7 2007		
10 15 07		

INVESTIGATING GROWTH TRAJECTORIES ON ENGLISH AS A SECOND LANGUAGE LISTENING AND READING COMPREHENSION TESTS

By

H. Gary Cook

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology and Special Education

2001

ABSTRACT

INVESTIGATING GROWTH TRAJECTORIES ON ENGLISH AS A SECOND LANGUAGE LISTENING AND READING COMPREHENSION TESTS

By

H. Gary Cook

American and Canadian university-based English language programs use a variety of assessments to evaluate their students. The *Standards for Educational and Psychological Testing* direct test developers and users to validate the assessments they create or use. A variety of validation studies have been conducted on university-based English language tests. Many studies have focused on the nature and components of language and language acquisition. Some studies have explored tests' general quality, but few studies have investigated how language tests function over time. This study explores the growth characteristics (trajectories) of Michigan State University's English language placement listening and reading comprehension tests. The focus of this research is to investigate the inferences made with these tests over time and to identify variables that may effect students' growth trajectories.

Subjects used in this study attended Michigan State University's English Language Center from 1992 to 1996. A total of 308 students coming from 28 different countries are sampled. Students' sex, age, academic intent, nationality, and length of stay at the English Language Center are obtained and used in statistical analyses. Four separate analyses are conducted. First, both listening and reading comprehension tests are evaluated using classical test statistical procedures. Second, tests are calibrated and equated using a Rasch item response theory model. Third, equated person ability estimates for the listening and reading tests are combined with demographic information and analyzed descriptively. Finally, a growth model study using a two level hierarchical linear modeling (HLM) analysis is conducted. Two HLM growth models are investigated: linear and quadratic.

Both examinations exhibited satisfactory classical and IRT test statistical characteristics. A linear based HLM model fit both the reading and listening comprehension tests best. On the listening comprehension test, students staying more than one term at the English Language Center have much lower growth trajectories. Older students have significantly lower initial test scores compared to younger students. Likewise, undergraduate students tend to have higher initial listening scores, and certain Asian language groups have significantly lower initial listening scores. On the reading test, older students staying more than one term have significantly higher growth trajectories, and Japanese and Middle Eastern students have significantly lower initial reading test scores.

This study provides a glimpse of growth trends on English as a second language listening and reading comprehension tests used for placement and achievement purposes at an American university. This study serves as a starting point in the investigation of growth on language tests. The study also provides a view into the differences between growth trajectories on listening and reading comprehension tests. This work introduces the notion of examining a test's growth trajectory to determine how different test uses might be appropriate or inappropriate.

This dissertation is dedicated to my father. He did not live to see the end of this work,
but the lessons he taught me on finishing what you start are responsible for its
completion. Thanks dad!

ACKNOWLEDGEMENTS

I am greatly indebted to several people, all of whom were instrumental in the inception and completion of this work. First, I would like to thank my committee for their patience and forbearance. I thank Besty Becker for coming in at the last minute and providing great editorial insight. Susan Gass has been both a great committee member and boss, and I thank her for her insights. I thank Ken Frank for his insistence that I get it right. His dogged editing and statistical insights greatly benefited this work. I am most grateful for Bill Mehrens' support and mentorship. To my committee, especially my chair, thank you! I would also like to thank Stephen Raudenbush for encouraging me to pursue this line of research for my dissertation.

As I trudged through this dissertation, my wife, Colet, has been the wind in my sails. This work would not have been completed without here loving support and understanding. I dearly love her and thank her. And given the chance, I'd marry her all over again. I thank my sons Cody and Alex, as well, for I have taken far too many evenings and weekends away from them. I am grateful to Ron Mahler, who urged me 17 years ago to consider going to college. I did, and this is the end of that journey.

Finally, I give glory and honor to Jesus Christ my Lord. I firmly believe that it is by the grace of God that I have been able to complete this work. I am humbled by his love and care for me. May I use the privileges and obligations that come with this degree for his honor and glory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background

> The Gileadites captured the fords of the Jordan leading to Ephraim, and whenever a survivor of Ephraim said, "Let me cross over," the men of Gilead asked him, "Are you an Ephraimite?" If he replied, "No," they said, "All right, say 'Shibboleth.'" If he said, "Sibboleth," because he could not pronounce the word correctly, they seized him and killed him at the fords of the Jordan. Forty-two thousand Ephraimites were killed at that time. (Judges 12:5-6)

In the book of Judges chapter 12, we read of a conflict between the Gileadites and the Ephraimites. Jephthah, leader of the Gileadites, was on a campaign against the Ammonites. He crossed by the country of the Ephraimites on that campaign. Once the Ammonites were defeated, he returned home. On his return trip, the Ephramites attacked claiming they were wronged by not being invited to join the campaign. Jephthah took up the battle, along with his kin from Gilead. The Ephraimites were routed. To ensure that Ephraimite warriors would all be dealt with, the men of Gilead set up boarder stations around Ephraim. Any man desiring entry into the land of Ephraim was given a test, *a language test*. All were asked to pronounce the word "shibboleth." The gloss in Hebrew is "a quick flowing stream." The men of Gilead correctly observed a dialectal difference between Ephraimites and other Israelites. Ephraimites did not produce the "sh" sound at the beginning of words. Instead of saying "shibboleth," they would pronounce the word as "sibboleth." Any man mispronouncing the word failed the test and was killed on the spot. The consequences of failing such a language test, the first recorded instance of a language test, were dire. From a student's perspective, it does not seem that the reputation of language tests have changed much.

Regardless of the somewhat dark beginnings, language testing has had a long and rich tradition of trying to properly assess a student's ability to communicate in another language. In recent history, language testing has functioned in many roles, chief being ascertaining an examinee's communicative competence in a second or foreign language. Since the turn of the century a wide variety of test measures have been created for this purpose both in Europe and North America (Barnwell, 1996). Recently, there has been a great rise in the use of *commercially* developed language tests, especially in the United States in the area of teaching English as a second language (TESL). Prior to the 1960s, most TESL programs developed their own exams. Today, programs using locally created exams are in the minority.

1.2 Common TESL Exams

The most prevalent commercially created TESL exams used in the United States (and in fact worldwide) are the Test of English as a Foreign Language (TOEFL) and the Michigan English Language Aptitude Battery (MELAB). The TOEFL is developed and administered by the Educational Testing Service and administered monthly throughout the world. This exam was originally created in the mid-1970s to screen incoming university students in the US and Canada whose native language was not English (Educational Testing Service, 1998). The MELAB is an exam offered by the University of Michigan. The MELAB is a second-generation test battery (first used in the early 1980s) developed by the University of Michigan for non-native English student screening. MELAB's predecessor, The Michigan Test of English Language Proficiency (MTELP), was the first widely used English language test battery in the United States. This test battery was created in the late 1950s. A large majority of TESL programs around the US and Canada uses one of these three tests.

The primary stated use for these exams is to establish English proficiency. Thus, universities (and other organizations, e.g., airlines and professional medical organizations) use these exams to ascertain an applicant's English ability prior to admission. If proficiency is adequate, applicants gain admission. If proficiency is inadequate, applicants are either not admitted or some sort of English remediation is required. The Educational Testing Service (ETS) and the University of Michigan (UM) state, in their technical manuals (Educational Testing Service, 1998 and University of Michigan, 1996), that these exams are to be interpreted in a norm-referenced fashion. That is, these exams are designed to compare examinees that take these tests with each other. This norm-referenced interpretation is further evidenced by results provided to examinees and institutions--College Entrance Examination Board Scores for TOEFL, and equipercentile equated scores for MELAB. Both organizations provide percentile rankings for scores in their publications explaining test use.

1.3 TESL Exam Purposes

While the stated use for exams like TOEFL, MELAB and MTELP is for proficiency purposes, many language programs use these measures for other types of decisions. In his texts on language testing and language program development, Brown (1995, 1996) suggests four main purposes for tests in second and/or foreign language education: proficiency, placement, diagnostic and achievement. Table 1 briefly outlines the goal of each of these test purposes.

Table 1: Description of Second/Foreign Language Test Purposes

Test Purpose	Goals of Test
Proficiency	To identify an examinee's general language proficiency by focusing on general language skills
Placement	To identify an examinee's level of language ability for the purpose of placing him or her into an appropriate level or class within a language program
Diagnostic	To identify an examinee's strengths and weaknesses in specific language skill areas within specific levels or courses
Achievement	To identify whether a student has mastered specific language skills taught within a language course

Decisions made with proficiency and placement tests are typically norm-referenced in nature. Diagnostic and achievement tests are given with mastery states in mind, and thus are criterion-referenced. As stated earlier, many language programs extend the purpose of commercially developed exams (TOEFL, MELAB, MTELP) to placement, diagnostic and achievement purposes as well. In fact, both ETS and UM offer retired versions of the

TOEFL and MTELP to language programs to use for placement and achievement purposes. In a note of caution, Brown (1995) writes,

> ...a test can be very effective in one situation with one particular group of students and be virtually useless in another situation or with another group of students. Teachers cannot simply go out and buy a test and automatically expect it to work with their students. It may have been developed for completely different types of students and for entirely different purposes. (p.119)

Crossing test purposes may limit the accuracy of decisions language programs make with externally created tests. Some researchers have argued that more language programs should create their own "in-house" exams (Brown, 1996; Cook, Dunsmore & Tan, 1998; Hughes, 1989). In-house exams would be directed at specific curricular goals and objectives within a language program and thus may provide better decisions. The quality of decisions made with in-house exams is dependent on the reliability and validity of the measures used. Unfortunately, many in-house examinations seldom have even basic reliability and validity studies.

Whether in-house tests or commercially purchased exams, such language tests are commonly used to measure students' achievement from term to term. Language programs use tests they create, adopt or adapt to determine whether a student has or has not completed--at an adequate level--curricular goals and objectives. Decisions made with these tests can, at times, have dire consequences for examinees. For example, passing a language test may qualify a student to begin academic work; likewise, failure to pass a test may prevent a student from entering a school or institution and require him or her to pay additional fees to upgrade his or her language ability. Thus, these exams take on the nature of high-stakes tests. Realizing the high-stakes nature of these exams, commercial test companies have conducted validation studies to support inferences made from their exams. However, much of the research on commercial and in-house exams focuses on "single-event" sampling. That is, tests are analyzed assuming that they are used to make one-time inferences. Students take exam "X" and inferences are made with the results. The strength of the inference or the nature of test use is then investigated. Yet validation studies and studies of test use seldom focus on longitudinal characteristics of language tests. That is, how do these tests function within examinees over time? If commercially purchased or in-house assessments are used to evaluate students' achievement over time, it seems reasonable that the characteristics of these tests should be investigated over time.

1.4 Major Research Questions and Study Overview

The English Language Center (ELC) at Michigan State University (MSU) is a service unit tasked with teaching incoming international students English. The ELC has been part of the MSU community for over 30 years. Over that entire period, the ELC has used in-house exams to assess its students. For the first 25 years, ELC test batteries were based upon the MTELP. The latest version of the ELC test battery (or placement test) is fashioned more like the TOEFL test--with some minor differences. The current version of the ELC placement test (ELCPT), which is now known as the MSU English Language Test, has four main subtests: grammar, writing, listening, and reading, and all portions of the test are given at the same sitting.

It has been curricular policy at the ELC to give students the opportunity to take the ELCPT at the end of every term. This policy has resulted in a relatively large corpus of student longitudinal data on ELC tests. As stated earlier, few studies have investigated the characteristics of longitudinal change in tests; thus, these data provide a unique opportunity to investigate how students' test scores change over time on a specific test within a specific institution (namely, MSU). Said differently, these data allow investigation of the validity of making inferences about student growth. That is the focus of the research reported here. Two subtests of the ELCPT are of particular interest here: listening and reading.

The questions addressed in this study are concerned with investigating the validity of inferences associated with the ELCPT listening and reading comprehension tests--particularly the inferences made about students' progress over time. The *Standards for Educational and Psychological Testing* (AERA, APA and NCME, 1999) provide specific guidance on the investigation of test purposes and test use. The opening chapter of the *Standards* defines validity as follows:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed scored interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated (p.9).

The *Standards* provide guidance and direction to test developers and test users on how to validate the inferences made with test scores. The following standards are germane to the study reported here.

Standard 1.1: A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of evidence and theory bearing the intended use or interpretation (p.17).

Standard 1.4: If a test is used in a way that has not been validated it is incumbent on the user to justify the new use, collecting new evidence if necessary (p.18).

Standard 9.1: Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences (p.97).

Standard 9.6: When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation (p.99).

Standard 13.2: In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose (p.145).

Standard 15.3: When change or gain scores are used, the definition of such scores should be made explicit, and their technical qualities should be reported (p.167).

The above standards speak to several issues investigated in this study. The ELCPT is used to make both placement and achievement interpretations regarding students' English language ability. Standards 1.1, 1.4, 13.2 state that these inferences and interpretations should be investigated, especially when the test is used for more than one purpose (e.g., placement and achievement decisions). Standards 9.1 and 9.6 indicate that the reliability and validity of test score inferences should be examined when different language groups take tests. Finally, Standard 15.3 speaks to investigating the technical quality of gain scores. While gain scores are not used per se, gains in students' ELCPT scores per term are used for achievement decisions. Thus it seems reasonable to examine the technical quality or for that matter the very nature of gain or gain scores on tests--specifically the ELCPT listening and reading comprehension tests. The specific issues addressed in this study can be encapsulated in the following questions:

1. What is the nature of student growth trajectories on the ELCPT's listening and reading comprehension subtests?
2. What demographic factors affect students' growth on the ELCPT's listening and reading comprehension tests?
3. What inferences might be made regarding the use of the ELCPT's listening and reading comprehension tests based upon discovered student growth trajectories and demographic factors?

The remaining chapters in this work address these questions. Chapter two provides a review of recent research on language testing, focusing on studies dealing with second language test validation and second language listening and reading comprehension test use. The second chapter also describes the rationale behind the study's methodology and research questions. The third chapter presents the study's methods, including a description of subjects, materials, procedures, and analyses. Chapter four presents results from classical and IRT test analyses as well as descriptive summaries of IRT ability estimates. Chapter five briefly introduces hierarchical linear growth models and presents results from the growth model analyses conducted in this study. Finally, chapter six discusses this study's major findings as well as this work's limitations, contributions and future directions.

CHAPTER 2: REVIEW OF THE LITERATURE

2.1 Overview

This chapter is divided into four sections. The first section outlines major validation studies of second and foreign language tests. These studies provide a framework for understanding why and how validation studies are conducted on second and foreign language tests. Following this, a description of skill-based language testing research is provided, narrowed specifically to listening and reading comprehension tests. The third section summarizes current "growth-based" research on language tests. Here growth-based research refers to studies investigating gain-based or longitudinal characteristics of language tests. The fourth section draws upon the research presented to generate this study's methodology and main research questions.

2.2 Second and/or Foreign Language Validity Studies

Language test validation studies have been conducted for a variety of purposes. A review of recent second and/or foreign language test validation literature reveals that the majority of reported studies address construct or criterion related validity. Most reported construct validity studies investigate the nature and components of language and language acquisition. Criterion-related studies typically are correlational in nature and compare tests of similar purpose to each other. Another group of validity studies combines content, construct and criterion related evidence to validate specific tests or test batteries. The following section reviews current research related to language test validation.

A foundational work on language test construct validation was conducted by John Oller. In Oller's (1979) text on developing and categorizing language tests, he introduced the notion of a unitary construct for language learning--what he called the unitary competence hypothesis (Oller, 1984). Prior to Oller's research, teaching and testing isolated language skills (grammar, listening, speaking, reading, or writing) was thought to be an appropriate method of instruction and assessment. Through factor analytic techniques on a variety of language tests, Oller argued that isolating individual skills for instruction or assessment was inappropriate.

> In spite of all the remaining uncertainties, it seems safe to suggest that the current practice of many ESL programs, textbooks, and curricula of separating listening, and reading and writing activities is probably not just pointless but in fact detrimental.... It would appear that every teacher in every area of the curriculum should be teaching all of the traditionally recognized language skills. (p. 457, 458)

Oller (1981) carried this notion further to suggest that language and intelligence were the same thing. This claim generated "considerable theoretical controversy and focused further research on the nature of language proficiency" (Harley, Cummins, Swain and Allen, 1990, p.9).

In a seminal counter argument to Oller's claim, Bachman and Palmer (1982) argued that language learning (proficiency) was not a unitary construct but multidimensional. Like Oller, Bachman and Palmer investigated several different types of language tests assessing a variety of skills; however, their findings indicated that language proficiency was not necessarily a unidimensional construct. In fact, they posited that the nature of language proficiency might in fact be multidimensional. Carroll (1983) suggested that language proficiency might be somewhere in between unidimensional and multidimensional. In a follow up study, Fouly, Bachman and Cziko (1990) found after conducting a confirmatory factor analysis on language tests (similar to Oller's and Bachman and Palmer's research) that both multidimensional and unidimensional models of language proficiency could be supported. Studies mentioned thus far focused primarily on the following types of language tests: listening comprehension tests, reading comprehension tests, cloze tests (for more on cloze tests see Hughs, 1989), oral interviews, etc. The main aim in all of the aforementioned validity studies was to examine the nature (construct) of language learning. During much of the late 1970s and early 1980s, language testing construct validity research was concerned with this aspect of validity.

Another line of test validation research began in the early 1980s. This research addressed the efficacy of using latent trait models (initially Rasch models) on language tests. In the very first edition of *Language Testing*, Perkins and Miller (1984) compared the differences in analyzing a reading test using classical test theory and item response theory (IRT). In their analysis they conclude,

> [T]his investigation has shown that the Rasch one-parameter latent trait model detected more misfitting or weak items from a multiple-choice reading comprehension test than did classical test theory indices. And most importantly, the Rasch Model allows us to define the latent trait which is manifested in observable test behavior and to determine what the variable seems to be (p. 31).

The next edition of *Language Testing* featured an article by Henning (1984) who argued for the efficacy of using latent trait models (specifically the Rasch model). Like Perkins and Miller (1984), Henning also analyzed a reading comprehension test. He concluded that the Rasch model "produced numerous advantages for the test developers" (p.132). McNamara (1990) is also noted for investigating the use of the Rasch model. He finds that IRT is a useful tool for investigating language tests' underlying constructs. He focused specifically on speaking and writing tests. However, researchers began questioning whether language tests could adequately meet the unidimensionality or local independence assumptions required by latent trait models (Blais and Laurier, 1995; Henning, 1989, 1992; and Choi and Bachman, 1992). No real consensus has been reached by language test researchers regarding the use of IRT on language tests. IRT is still a commonly used model for scaling and evaluating language tests by commercial language test vendors and university language programs. There has, however, been a heightened awareness of the need to check assumptions when analyzing language tests with latent trait models.

Few reported studies have been conducted comparing one language test with another (criterion-related validity studies), especially commercially developed tests. One notable explosive exception is Bachman, Kunnan, Vanniarjan and Lynch (1988). As stated earlier, U.S. and Canadian private, college and university TESL programs commonly use one three commercial language tests TOEFL, MELAB or MTELP. However in Great Britain, a variety of examinations are used (Davies and West, 1989). In the U.S. and Canada, TESL programs most commonly use TOEFL. In Great Britain, British private and university language programs most often use the Certificate for

Proficiency in English (CPE). The British Council/University of Cambridge Local Examinations Syndicate (UCLES) develops the CPE, and the Educational Testing Service (ETS) develops TOEFL.

In the Bachman, et al. (1988) study, the TOEFL and CPE were compared on both content and construct comparability. The study found both similarities and differences between the two examinations but ended by stating that differences between the two exams were not great. This finding generated a visceral response from Henning (1989), who questioned Bachman and his colleagues' credibility, since they were funded by UCLES to conduct the study. Bachman's (1989) rejoinder in the same journal edition countered that Henning (an employee of ETS) was himself biased. Bachman and colleagues have since published a text describing the entire comparability study (Bachman, Davidson, Ryan, and Choi, 1994). To date, no follow-up study has compared commercially developed proficiency examinations.

Another class of validity studies looks at specific language tests or test batteries and evaluates validity and reliability. Few of these studies are mentioned in the literature. The first recorded study of this kind was conducted by Davies (1984). In his study Davies described the validation of three English language proficiency tests: the English Proficiency Test Battery, the English Language Battery, and the English Language Testing Service. The general aim of Davies' research was to show how validation studies on specific tests might be conducted. Davies discussed methods of item and reliability analysis. He also discussed the use of factor analysis to investigate test construct validity. Davies described several criterion-related validation procedures: correlating test scores with student grades, test scores with teacher perceptions, test scores with longitudinal academic performance, and test scores with teacher and student questionnaires. A brief description of setting cut-off scores was mentioned as well. Davies concluded by saying, "Test validation is a necessary part of test construction.... The process of concurrent and predictive validation, the internal analyses and external comparisons are time consuming but routine" (p.68).

Ten years after Davies' study, Wall, Clapham and Alderson (1994) presented a study evaluating a language program's placement test. The Language Testing Research Group at Lancaster University was asked to evaluate the Institute for English Language Education's placement test at Lancaster University. Wall, et al. write, "The nature and validation of placement tests is rarely discussed in the language testing literature, yet placement tests are probably one of the commonest forms of tests used within institutions.... (p.321). Wall, et al. addressed four major questions about the investigated test: 1) does the placement test correctly identify students' language needs?, 2) do students feel they are appropriately measured?, 3) is the content of the test appropriate?, and 4) is the test reliable? Through extensive questionnaires given to teachers and tutors, they concluded that the test was meeting students' needs. Through student questionnaires they determined that students felt adequately assessed. Wall, et al. reviewed the content of the test and interviewed one instructor and had mixed reviews of the placement test. They felt that the format of the exam may have been sending pedagogically inappropriate messages to students. Reliability coefficients for the exam were presented and were moderately high (>0.75). While the test evaluation was considered acceptable, Wall, et al. expressed concerns about their findings.

> What this study has demonstrated, we believe, is the difficulty of validating a placement test against external and possibly even internal criteria. We feel that this important matter has been neglected, and believe more discussion is needed in the language-testing literature about the most appropriate ways in which a placement test can be validated (p.343).

Fulcher (1997) conducted the most recent and by far most thorough placement test validation study. Fulcher evaluated the English Language Institute's placement test which was used to place students into the University of Surrey's English for Academic Purposes program. The study utilized a variety of measurement techniques to evaluate the test. For item analysis, a Rasch IRT model was used. The author described misfit analysis using obtained Rasch estimates. Factor analytic techniques were used to investigate the construct validity of the test battery. Both concurrent and predictive validation procedures were used to identify the tests' criterion-related validity. Fulcher reported KR20 reliability estimates, which were somewhat low due to the short number of items on the test, and an evaluation on the method of setting the cut-score for the examination was described. The author found that generally the test was providing adequate information, but he commented that the reading portion of the test needed to be lengthened to provide more reliable information. In a section titled "Future research," Fulcher made the following comment.

> It is a truism that in British universities there is no evidence to suggest that after a certain amount of time on any given programme, a student will 'improve' by 'x' whatever the unit 'x' is in terms of ability. Indeed, score gain studies, even if conducted, would be meaningless, without a clear understanding of what 'x'

> is....As such, score gain studies may be conducted with new groups of students over different timescales, following different courses. If successful, this would allow the English Language Institute to say to a department that a particular student would require (given error) from 'a' to 'b' months of language tuition [course-work] to reach a level at which he or she would, with *p* probability, be able to cope with an academic course with (or without) English language support.
>
> This kind of information is not currently available, but in principle could be, as the result of a careful development of research within the context of specific language programmes of large educational institutions (p.137).

To summarize this section, language test validation studies have been evolving over that last 20 years. Early research focused on the nature of language learning and found that both unitary and multidimensional models could describe test data. The overwhelming acceptance of latent trait models, as evidenced by Henning (1984), Perkins and Miller (1984) and McNamara (1990), has been tempered by research showing strong multidimensional characteristics of language tests (Blais and Laurier, 1995; Henning, 1989, 1992; and Choi and Bachman, 1992). Few studies compare commonly used commercially available language tests for criterion-related validity purposes, and the studies that have been presented are highly contentious. This has moved the field to begin investigating institutionally created examinations. Both proficiency examinations (Davies, 1984) and placement tests (Wall, et al, 1994 and Fulcher, 1997) have been evaluated. Studies of institutionally created examinations have provided guidelines for useful evaluative techniques, e.g., IRT analysis, factor analytic procedures, criterion-related validity studies which include teacher and student information, content review, comprehensive reliability analysis, and evaluation of cut-score creation and use. However, findings from the evaluation of institutionally created tests have created several questions. A call for more appropriate ways to validate tests has been made (Wall, et al. 1994). Further, a call for gain studies has been made (Fulcher, 1997). The results of gain studies could assist TESL program administrators in making better inferences with test scores and better decisions about students' English language abilities. As Fulcher (1997) writes, "The goal of placement testing is to reduce to an absolute minimum the number of students who may face problems or even fail their academic degrees because of poor language ability or study skills" (p.113).

2.3 Second Language Listening and Reading Test Studies

Listening and reading comprehension tests are common subskill tests used in second and foreign language proficiency and placement examinations. At U.S. and Canadian colleges and universities, it is often necessary for non-native English students to take notes and listen to lectures. The ability to take notes and follow a classroom lecture is crucial to success at university and requires sophisticated listening skills. Listening tests are designed to evaluate these skills. Commonly, listening tests are multiple-choice and assess note-taking skills, short-term recall of informal, academic conversations, and short lectures. Equally, effective reading skills are a key to success at university. Undergraduate and graduate students must read pages-upon-pages of text. It is not uncommon for students to have 100 pages of required reading per night for homework. Students without the necessary skills to effectively read this material are greatly disadvantaged, if not lost. English as a second language reading comprehension tests are designed to evaluate whether students have basic reading skills necessary to sustain success in school. Reading comprehension tests are almost always multiple-choice and typically assess skimming and scanning skills, vocabulary knowledge, and comprehension skills. Dunkel (1993) reports on a survey asking college professors which skills they thought were most important for foreign language students they were teaching. She writes,

> When asked to indicate the relative importance of listening, reading, speaking and writing for international students' success in their academic departments, U.S. and Canadian professors of engineering, psychology, chemistry, computer science, English and business, for example, gave the receptive skills of listening and reading the highest ratings. (Reading comprehension was seen as most important of the four abilities in all disciplines surveyed except English.) Listening was rated the second most important in four of the six disciplines (engineering, psychology, chemistry, and computer science) (p. 267).

Because of the important nature of listening and reading comprehension tests, a number of studies have been reported on them in the language testing literature.

de Jong (1984) provided a glimpse of a Dutch pre-university English listening comprehension test for the Dutch National Institute for Educational Measurement. The pre-university listening test had two item formats (true/false and modified cloze), and the psychometric properties of item formats were the focus of the study. His findings indicated that the modified cloze task had better psychometric qualities.

Researchers and test users are often interested in what affects listening comprehension test performance. Two studies particularly focus on this issue. Hale and

Courtney (1994) were interested to know if note-taking affected performance on the TOEFL short monologues or minitalks. On the TOEFL, one section of the listening subtest requires students to listen to short dialog and respond to questions. During administration of this section, examinees are not allowed to take notes. Hale and Courtney wanted to know if it mattered. It did not. Apparently the cognitive load required while taking notes during a short monolog or minitalk overwhelmed examinees to the point where taking notes could even be seen as a detriment. Jensen and Hansen (1995) studied whether prior-knowledge of a subject prebiased non-native English speakers who knew the topic being tested. On the University of Kansas' Test of Listening for Academic Purposes, they found that, indeed, there were significant differences in the performance of students who had prior knowledge. However, the actual differences between students who had prior knowledge and those who did not was so minor that differences were considered to be practically non-significant. Freedle and Kostin (1999) studied whether minitalk (short monologues) passages on the TOEFL had any effect on test performance. Researchers claimed that students taking listening and reading comprehension tests did not need the passages to answer questions. In essence, the claim was made that the questions could stand alone; hence, comprehension tests were more vocabulary or item level tests. Kostin and Freedle found that substantial variation in test performance was accounted for by minitalk passages. They found that having passages does count.

Buck (1990, 1991) conducted his doctoral work investigating the introspection of students taking listening comprehension tests. In a qualitative/quantitative study Buck administered a listening comprehension test to six students. The test had 13 sections, and after each section, students answered questions. Once complete, interviewers surveyed students asking why they responded as they did. Several interesting issues involving test format and students' perception of what formats were intending to evaluate were discovered. In his conclusion, Buck argues that listening comprehension tests seem to assess more than listening. Cognitive skills seemed to also be assessed by these examinations. Essentially, Buck argued that listening comprehension tests were inherently multidimensional. However, McNamara (1991) made exactly the opposite claim. McNamara studied the listening portion of the Occupational English Test given by the Australian National Office of Overseas Skills Recognition. The focus of his research was whether a Rasch model was appropriate for this exam--specifically the issue of test dimensionality. McNamara writes,

> The controversial issue of test dimensionality in IRT analyses has been investigated in this paper using both theoretical and empirical approaches. The conclusion reached in each case is that the misgivings sometimes voiced about the limitations, or indeed, the inappropriateness of Rasch IRT for the analysis of language test data may not be justified.

Early research on reading comprehension tests focused on several issues; one line of research dealt with the use of cloze tests. Oller (1979) championed integrated tests (tests requiring more than one language skill) and felt they were superior to multiple-choice discrete-point tests. The cloze test was thought to be a holistic assessment of reading ability. Several approaches to researching the cloze test have been presented. Studies have compared performance on cloze tests with performance on multiple-choice reading comprehension tests (Bensoussan, 1984 and Hale, Stansfield, Rock, Kicks, Butler and Oller, 1989). Variants of the cloze procedure have been created and several studies have investigated the relationships between cloze tests and their variants (Chappelle and Abraham, 1990; Farhady and Keramati, 1996; and Klein-Braley, 1997). In the realm of construct validity, Turner (1989) investigated the underlying trait structure of the cloze procedure. While the cloze procedure/test is an interesting method for evaluating reading, its use has not been widespread on commercial language tests (either TOEFL or MELAB). Hence, many language programs do not use cloze tests or use them in conjunction with a multiple-choice reading test. More common to proficiency and placement test batteries is the multiple-choice reading comprehension test.

Several studies have been conducted on reading comprehension tests. In an early study Shohamy (1984) investigated whether test method, i.e., multiple-choice vs. open-ended, affected test performance. She found that it did. She also found that students tended to perform better on multiple-choice and open-ended tasks when the questions were in their native language. Freedle and Kostin (1993) investigated whether item difficulty on TOEFL reading comprehension items could be predicted based on the linguistic and textual features of the passage and test item. They found that "a significant amount of item difficulty variance can be accounted for by a relatively small number of variables for the three reading item types studied (main ideas, inferences and supporting ideas)" (p.167). Lumley (1993) studied whether TESL teachers were able to identify the difficulty of subskills on a reading comprehension test as well as ranking what the difficulty of specific items might be. Using an English for academic purposes test developed for the University of Melbourne, Lumely asked 5 teachers to isolate the subskills on 22 of the 58 reading comprehension test items and rank them for difficulty.

A follow up task required teachers to give their rankings on the difficulty of specific items. Using a Rasch IRT model to calibrate the test, Lumley compared the teachers' rankings of the subskills and items. He concluded that the findings "lend some empirical support to the value of using teachers' judgements in examining test content,.... The judgements they make about linguistic matters in test design and content validity also have significance for teaching" (p.231). In a construct validation procedure, Anderson, Bachman, Perkins and Cohen (1991) examined the construct validity of the Educational Testing Service's Descriptive Test of Language Skills--Reading Comprehension Test. Anderson, et al. used three techniques to explore the test validity: think aloud-protocols, test item content analysis, and test performance. Their article concluded as follows.

> Perhaps the greatest insight gained from this investigation is that more than one source of data needs to be used in determining the success of reading comprehension test items. By combining sources of data...greater insights are gained into the reading comprehension process as well as the test taking process. This information is valuable for test developers in evaluating test items, as well as for classroom teachers... (p.61).

To summarize, a variety of studies have been conducted in the language testing literature on listening and reading comprehension tests. Test formats have been investigated (de Jong, 1984; Freedle and Kostin, 1999, Oller, 1979, Shohamy, 1984). Researchers have investigated what affects test performance on listening and reading comprehension tests (Hale and Courtney, 1994; Jensen and Hensen, 1995). Test difficulty or the predictability of test difficulty has been studied (Freedle and Kostin, 1993; Lumely, 1993). The use of IRT as a testing methodology on listening and reading comprehension tests (de Jong, 1984; McNamara, 1991) and the construct validity of listening and reading comprehension tests has been examined (Buck, 1990, 1991; Anderson, et al. 1991). However, one area of study receiving little attention is how actual listening and reading comprehension test results are used to make decisions about students' abilities. As stated earlier, Wall, et al. (1994) challenged researchers to look for new (or more appropriate) ways of validating language tests. A much more specific challenge has been raised by Fulcher (1997)--the issue of studying gains. Fulcher argued that with the advent of IRT and the ability to have "test-free" estimates of students' abilities, predictability studies could be conducted providing university departments with estimates of how long and in what manner students might finish their language requirements. Unfortunately as Fulcher surmised, few studies have been conducted on language tests focusing on how student scores change over time. The next section describes the few research studies that deal, albeit indirectly, with this issue.

2.4 Growth-based Research Studies

Henning (1982) conducted one of the earliest growth-based studies in second or foreign language testing. Recognizing that information about gain or growth could greatly aid in "program intervention and reform" (p.467), Henning designed a procedure allowing for growth-referenced evaluation of language tests. This "technique was devised for the measurement and comparison of rates of learning in six component skills of English language proficiency" (p.473). Henning used the Ain Shams University (Egypt) English Proficiency Exam for his analysis. Using subtest reliability estimates, inter-test and inter-item correlation coefficients and correlations between students' year of study and total test performance, Henning created a variety of indices (importance of component skills index, program learning rate, commensurate growth rate) which culminated in the critical intervention index (CII). The CII was claimed to be a measure of the relationship of the "priorities for instructional intervention" (473). All indices were said to be similar to correlation coefficients in that they were bounded between minus one and plus one. Henning concludes his study with the following comment.

> The prospect of evaluative methodology that is sensitive to comparative rate of achievement gain rather than absolute gain, and which promises to pinpoint program weaknesses for reallocation of resources in a less subjective manner warrants still more vigorous investigation (p.476).

In his study, he urged other language testing specialists to utilize his measure of "growth-referenced evaluation" in other educational contexts. To date, no follow up study using Henning's procedure has been reported. The lack of studies using this procedure should not indicate a lack of concern over monitoring and measuring gain over time in language tests. Henning's procedure was somewhat cumbersome and psychometrically questionable. Also, until recently no effective statistical technique had been able to adequately capture or model gain or growth over time.

No additional studies have been reported in the second or foreign language research literature focusing on test performance over time. However, two studies have been reported focusing on modeling growth or change over time. The first of these studies is Pienemann, Johnston and Brindley (1988). Pienemann and his colleagues (Meisel, Clausen and Pienemann, 1981 and Pienemann and Johnston, 1987) have focused on the acquisition order of German and English. Pienemann, et al. presented a predictive framework for the stages of language acquisition. The research reported in this study focused primarily on syntactic acquisition of English. Using this framework, Pienemann, et al. created a profile analysis assessment, which identified the developmental stages of examinees. In the conclusion, the authors warned that,

> While it may be tempting for educational administrators...to use the results of a quantifiable language test to stream learners into classes or to just justify funding decisions, it is important to point out that the present assessment procedure was not designed for these purposes (p.240).

The authors asserted that this procedure was meant to be used for pedagogical purposes not programmatic decisions. Conceptually, Pienemann, et al. created a model of language acquisition and created a measure to evaluate that model. As an aside, the profile analysis mentioned in this article has continued in its development, but no validation studies of this assessment have been reported.

The second study (Mellow, Reeder and Forster, 1996) investigated the use of time-series analysis to research developmental change. Mellow, et al. sought to monitor students though their developmental courses (stages of acquisition). The results from two studies were reported. Each study investigated pedagogical interventions and used time-series analysis to monitor effects. The stated goal of Mellow, et al. was to showcase time-series analysis as a possible option for second language researchers to evaluate longitudinal data.

As seen by the research briefly summarized above, the number of studies investigating change over time on language tests is sparse--at best. Few studies of change have been reported in the language testing literature, yet language programs routinely use proficiency and placement tests to monitor and determine student achievement over time. Since little is known about the nature of growth on language tests, are the inferences we are making with the tests we use reasonable? Are we making the best decisions? What influences performance over time on language tests and what effect might those influences have on inferences made about students' placement or advancement? The purpose of the research presented here is to begin investigating these questions.

2.5 Research Methodology and Background Research for Research Questions

Davies (1984), Wall, et al. (1994), and Fulcher (1997) have all outlined procedures for evaluating proficiency or placement tests. Were one to synthesize the methodology for evaluating proficiency or placement tests described in these studies, the following set of procedures would emerge:

1. conduct a classical item analysis and reliability study,
2. calibrate the test(s) using latent trait models (Rasch is most often used),
3. conduct a misfit analysis to evaluate latent trait model fit,
4. conduct validity studies--to include but not limited to content analysis, criterion-related validity analysis and a construct validity study, and
5. evaluate the cut-scores or decision(s) made with the test(s).

Similar types of procedures can be found and recommended in graduate level measurement texts (e.g., ; Allen and Yen, 1979; Brown, 1996; Crocker and Algina, 1986, and Henning, 1987). Many, if not most, techniques for proper test analysis are commonly understood and agreed upon. When investigating the validity and reliability of second or foreign language proficiency or placement tests, one would be well served to follow the evaluative methodology outlined above.

Unfortunately, little guidance can be found on the investigation of gain scores or growth characteristics on language tests over time in the language testing literature. Before the mid 1980s, it would be difficult to find much research investigating the role of change over time in general educational settings. Fortunately techniques for investigating longitudinally based data have been developed. These techniques have been referred to in a variety of ways: multi-level linear models, random-effects models, and hierarchical linear models (Bryk and Raudenbush, 1992, p.3). For this paper, these techniques are referred to as hierarchical linear models (HLM). The power of HLM lies in the ability to model growth on a variety of nested educational data. In educational settings, these growth models have been used to investigate everything from children's vocabulary development (Huttenlocher, Haight, Bryk, and Selzter, 1991) to estimating school effectiveness (Willms and Raudenbush, 1987). Armed with HLM, researchers now have the ability to address Fulcher's (1997) challenge to begin studying the effects of gain on language tests. Using this research methodology, it is possible to investigate what inferences might be made with language tests as they are used to monitor student language achievement over time. It is this and previously mentioned research that motivates the study presented here.

However, prior to presenting the research questions, one more area of research should be briefly outlined--what we know about language learning. In an article "staking out the territory" of second language acquisition, Larsen-Freeman (1993, 154-156) offers the following ten general characteristics of language learning that second language educators should be aware of:

1. the learning/acquisition process is complex,
2. the process is gradual,
3. the process is non-linear,
4. the process is dynamic,
5. learners learn when they are ready to do so,
6. learners rely on the knowledge and experience they have,
7. it is not clear from research findings what the role of negative evidence is in helping learners to reject erroneous hypotheses that they are currently entertaining,
8. For most adult learners, complete mastery of L2 may not be impossible,
9. there is tremendous individual variation among language learners,
10. learning a language is a social phenomenon.

Larsen-Freeman mentions that language learning is a non-linear process. What type of non-linear process is it? Can this process be seen in performance on listening and reading comprehension test scores? These inquiries motivate this study's first research question: What is the nature of growth trajectories on the ELCPT's listening and reading comprehension subtests?

To investigating growth trajectories, it is necessary to conduct proper test analyses. These analyses include classical test analysis and IRT test calibration. One of the generated outputs of most common IRT analyses is person-ability estimates or scores. Person ability scores are valuable since "the ability θ [ability score] of an examinee is monotonically related to the examinee's true score..." (Hambleton, Swaminathan and Rogers, 1991, p.77). The use of person ability scores will enable growth models to be estimated, since each obtained estimate is a representation of an examinees' true score at the time of testing.

Larsen-Freeman (1993) argues that there is substantial individual learner variation in language acquisition. Of particular interest here are easily collected unique demographic learner variables such as sex, age, language group, and academic intent for studying English. These issues are the focus of the second research question: What demographic factors affect students' growth on ELCPT's listening and reading comprehension tests?

The last research question investigates the likely inferences that can be made on these tests based upon the discovered growth trajectories and demographic variable influences. Prior studies on placement tests (Davies, 1984; Wall, et al., 1994; and Fulcher, 1997) focused on portraying methods of evaluating those types of tests. They did not address how decisions were made with the tests they investigated. Similarly, listening and reading test studies examined issues related to test format and predicting test difficulty. The use of IRT and the characteristics that affect test performance have been studied as well. But little research has addressed the efficacy of inferences made with language exams over time. The third research question addresses these issues: What inferences might be made regarding the use of the ELCPT's listening and reading comprehension tests based upon discovered student growth trajectories and demographic factors?

CHAPTER 3: METHODS

3.1 Subjects

The subjects used for this study are international students who have taken Michigan State University's, English Language Center (ELC) Placement Test between fall 1992 and fall 1996. To be included in this sample, students must have had at least two recorded test scores in their records and had at least their sex identified on their individual records. Response data were collected for both the listening and reading subtests. A total of 308 students are used for this study, but only 121 students (39.3%) took both listening and reading tests. There are 199 students in the listening analysis and 228 for the reading analysis. Table 2 below presents the numbers of students by sex and subtest.

Table 2: Frequency of Subject's Sex by Subtest

Sex	Listening Test	Reading Test
Male	116 (58%)	126 (55%)
Female	83 (42%)	102 (45%)
Total	199	228

Over the period sampled for this study, there were slightly more male students than female.

Students used in this study come from a variety of language backgrounds. On students' individual record forms, they self-report their native language and country of residence. Students in this sample came from 28 different countries. Many countries in this sample had only one student represented. Several countries represented share a common language; for example China, Taiwan, and Hong Kong residents speak

Chinese. Further, several countries represented in this sample share a common language group; for example Saudi Arabia, Kuwait, and Qatar residents speak Arabic. Because of inadequate sampling of some countries and the possibility of aggregating students together, students are broken into 5 common groups: Chinese (CHI), Japanese (JPN), Korean (KRN), Mid Eastern (ME), and Other (OTH). The Other group represents students from a variety of languages and countries. Table 3 presents the frequency count of students' country by language group.

Table 3a: Country by Language Group for Listening Subtest

Country	OTH	ME	CHI	JPN	KRN	Total
Brazil	1					1
China			2			2
Colombia	1					1
Cyprus	1					1
Egypt		1				1
Germany	1					1
Hong Kong			3			3
Indonesia	1					1
Japan				42		42
Jordan		2				2
Korea					90	90
Kuwait		2				2
Kazhakstan	1					1
Malaysia	3					3
Qatar		3				3
Russia	3					3
Saudi Arabia		8				8
UAE		2				2
Thailand	9					9
Tunisia		1				1
Turkey	3					3
Taiwan			18			18
Venezuela	1					1
Total	25	19	23	42	90	199

Table 3b: Country by Language Group for Reading Subtest

Country	OTH	ME	CHI	JPN	KRN	Total
Argentina	3					3
Brazil	1					1
China			18			18
Finland	1					1
Germany	1					1
Greece	1					1
Italy	1					1
Japan				46		46
Jordan		1				1
Korea					97	97
Malaysia	1					1
Missing	14					14
Saudi Arabia		16				16
Taiwan			14			14
Thailand	8					8
Turkey	4					4
Vietnam	1					1
Total	36	17	32	46	97	228

Note that 14 students in the reading test sample (Table 3b) did not have their language information available. They are grouped into the "Other" category. Table 4 shows the breakdown of language group by subtest.

Table 4: Language Group by Subtest

Language Group	Listening Test	Reading Test
Other	25 (13%)	36 (16%)
Middle Eastern	19 (10%)	17 (7%)
Chinese	23 (12%)	32 (14%)
Japanese	42 (21%)	46 (20%)
Korean	90 (45%)	97 (43%)
Total	199	228

Both the listening and reading subtests have similar distributions of language groups.

Korean, Japanese and Chinese students make up the bulk of subjects (77%). The subjects collected in this sample do not seem to deviate from normal international student enrollment at MSU over the period of this study (Office of International Education Exchange, 1993, 1997).

Table 5 presents descriptive statistics for students' age by language group and by subtest.

Table 5: Descriptive Statistics of Student Age by Language Group by Subtest

Language	Listening Test					Reading Test				
Group	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
Other	24.14	6.17	19	46	25	25.69	6.73	18	48	36
Middle Eastern	27.05	5.15	17	34	19	27.12	5.11	19	36	17
Chinese	24.61	4.86	18	43	23	25.09	3.05	19	34	32
Japanese	22.95	3.53	18	35	42	23.37	4.74	19	44	46
Korean	23.28	3.48	18	41	90	23.76	4.34	18	43	94
Total	23.87	4.35	17	46	199	24.43	4.88	18	48	225

Student ages across language groups are similar with a few exceptions. On both subtests, Middle Eastern students tend to be the oldest by almost two years. Japanese students, on average, tend to be the youngest. Note that the reading subtest N for students is 225; three Korean students in this sample were missing their ages.

International students come to MSU for three primary reasons: to study English (either for enrichment or preparation for university coursework), to study in an undergraduate program, or to study in a graduate program. Table 6 presents the academic status of the students sampled in this study. Academic status is identified by students' responses on the ELC application form.

Table 6: Frequency of Student Academic Status by Subtest

Academic Intent	Listening Test	Reading Test
Studying English	97 (49%)	86 (38%)
Undergraduate Student	64 (32%)	89 (39%)
Graduate Student	38 (19%)	49 (21%)
Missing	--	4 (2%)
Total	199	228

The listening test sample has a higher percentage of students studying English compared to the reading test sample; conversely, the listening test sample has slightly lower percentages of undergraduate and graduate students.

Table 7: Descriptive Statistics of Student Age by Academic Status by Subtest

Academic Status	Listening Test					Reading Test				
	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
Studying English	23.31	3.35	17	35	97	24.15	4.56	18	44	86
Undergraduate Student	22.17	2.86	18	30	64	22.46	3.28	18	36	89
Graduate Student	28.13	5.86	20	46	38	28.57	5.37	20	48	49
Missing	---	---	---	---	---	23.50	4.04	20	29	4
Total	23.87	4.35	17	46	199	24.43	4.88	18	48	228

For both the listening and reading subtests, undergraduate students tend to be the youngest, and graduate students the oldest (from Table 7). This is not an unexpected result. Graduate students have finished 4 years of schooling prior to arriving at MSU. There is a mixture of ages for students studying English. This mixture includes students studying English for enrichment, students preparing to enter undergraduate school, and students preparing to enter graduate school.

3.2 Materials: English Language Center Listening and Reading Comprehension Tests

The tests used for this study are the ELC Placement Test Battery (ELCPT) listening and reading subtests. These tests are used for initial placement and end of term progress. International students seeking admission to MSU may provide results from the TOEFL or MELAB exams to gain entrance. However, students not meeting MSU's English language minimum requirement are required to take the ELCPT.

The Listening test is a standard English as a second language (ESL) multiple-choice listening comprehension test. It has four main sections. Section one is a short lecture (7-10 minutes) on a non-academic topic. Section two has several short conversations where students listen and answer questions about each conversation. Section three has several short listening passages; students listen and answer questions about each passage. The last section presents questions related to the lecture heard in section one. The test is designed to ascertain a student's ability to comprehend basic English language interactions experienced at an American university. In total the listening test takes approximately 35 minutes to complete. Students' responses to questions are placed on answer sheets and scored electronically. There are two versions of the listening test: L92-1 and L92-2. Each version has two forms (A & B). Dual forms are used to limit the amount of cheating during actual test administration. Both examinations came into service in 1992.

The Reading test is similar to most English as a second language (ESL) multiple-choice reading comprehension tests. There are three main sections. Section one has several short reading passages. Students read passages and answer questions. Section two displays a graph, figure or table. Students review the graph, figure or table and answer questions about it. The last section tests vocabulary. In this section, students either select an appropriate synonym or select a correct definition. This test is designed to assess basic English language reading skills needed at American universities. Student responses are recorded on answer sheets and scored electronically. There are 4 versions of the reading test: R91-1, R91-2, R91-3, and R92-1. Like the listening test, the reading test has dual forms to limit cheating. These examinations came into service at the ELC in 1991 and 1992.

Table 8 displays a breakdown of each subsection of each test.

Table 8 Subsections for ELC Listening and Reading Comprehension Tests

	Subsection			
Tests	1	2	3	4
Listening Comprehension	Brief Lecture	Short listening test items	Conversations test items	Lecture test items
Reading Comprehension	Short reading comprehension test items	Reading graph/ figure/ table test items	Vocabulary test items	

3.3 Procedures

Data collected for this study came from three sources: student personal records, the ELC central database, and the ELC testing office. These data included student identification number, sex, age, native language, academic status, listening and reading test scores, and term of test scores. All test answer sheets for listening and reading tests administered between 1992 and 1996 were assembled, and an IRT calibration of test scores was conducted. All ELCPT reading tests have common linking items to facilitate equating. Once these tests were calibrated all estimates were equated and placed on a common scale. The listening tests do not have common items. To equate listening test estimates, a common person equating procedure was used. Once equated, individual IRT person ability estimates were obtained for all students in the sample, and person ability estimates were matched with student identification numbers and terms in which scores were obtained. Students having two or more recorded scores were retained, and students with only one recorded score were removed from the sample.

After initial matching, student records were reviewed and each student's sex, age, language group, and academic status were added to a common file. An additional dummy variable was added after initial analysis ("More than one term"). It became clear that a majority of students had only two time points. This indicated that most students completed their ELC course requirement(s) for listening or reading in one term. Thus this variable was added to determine if students staying at the ELC for more than one term had different growth trajectories than students who stayed only one term. Information from the common file was used to generate descriptive statistics. To facilitate a hierarchical linear modeling procedure, the common file was split into two separate files: a level 1 file and a level 2 file. Table 9 displays the components of each file for the hierarchical linear modeling procedure.

Table 9 Level One and Two Variables for HLM Analysis

Level	Variables
Level One Listening File	Listening test weighted ability estimate per term academic term of test (e.g., 0, 1, 2, etc) academic term-squared
Level One Reading File	Reading test weighted ability estimate per term academic term of test (e.g., 0, 1, 2, etc) academic term-squared
Level Two Listening and Reading Files	Sex of subject Difference from mean age Academic intent (i.e., studying English, undergraduate, etc.) Language group of subject More than one term

In most IRT programs, item difficulties (b_i) and person ability estimates (θ_i) have associated standard error estimates (SE). When creating HLM files, person ability standard error estimates are used to condition or weight the outcome variables using the weight 1/ SE. Essentially, this weight represents an information function or the amount of information (accuracy) a particular person ability estimate is predicted to have. Weighting by the information function will provide a more accurate view of how students' abilities are changing over time.

3.4 Analyses

Four unique sets of analytic procedures are used for this study. First, listening and reading test data are analyzed using classical test theory methods with ITEMAN (Assessment Systems Corporation, 1995a). Each test's descriptive characteristics and reliability are investigated.

Second, Rasch model item response theory (IRT) analyses is conducted using RASCAL (Assessment Systems Corporation, 1995b). The Rasch IRT model is used primarily because of small sample sizes and ease of equating (Lord, 1983, Hambleton and Swaminathan, 1989). Rasch models are also commonly used in the analysis of language tests (Henning, 1984; Perkins and Miller, 1984, McNamara, 1990). When using IRT calibration, RASCAL requires researchers to choose a method of centering. For this study all tests will be centered on item difficulties. The RASCAL program allows users to save person ability as well as item difficulty estimates. To ensure that all person ability estimates are equivalent, test equating procedures are conducted. As stated earlier, all ELCPT reading tests have common items. Person ability estimates for reading tests are equated using the Wright and Stone (1979) equating procedure for common item equating. Listening tests are equated using common person equating. The procedure used for common person equating is a variant of the common item equating method mentioned in Wright and Stone (1979).

Third, equated person ability estimates for the listening and reading tests are combined with aforementioned demographic information and analyzed descriptively using SPSS 7.5.1 (SPSS. Inc., 1996). A variety of descriptive and graphic analyses are conducted with the purpose of investigating trends in the data. Prior to converting files into HLM format, all term variables are centered in SPSS.

For the last analysis, a growth model study using a two level hierarchical linear modeling (HLM) analysis is conducted (Bryk and Raudenbush, 1992). Two HLM growth models are investigated: linear and quadratic. The standard linear growth model is used as a starting point (Bryk & Raudenbush, 1992, Chapter 6). Larsen-Freeman (1993), in an article summarizing major findings in the field of second language acquisition, suggests that language growth is gradual and a "non-linear process" (p.154). Long (1990) directly suggests that language acquisition is in fact curvilinear. Cook (1995) investigated language tests using a quadratic HLM model and found a high degree of fit. Thus, the use of a quadratic model in the analysis seems justified.

CHAPTER 4: TEST ANALYSES AND DESCRIPTIVE SUMMARIES OF ABILITY ESTIMATE DATA

4.1 Classical Test Theory Results

The first analysis conducted is an investigation into the classical test theory characteristics of the listening and reading tests. Note that the sample sizes presented for classical test statistics are much greater than the sample sizes used in most of this study. For the classical test analysis and IRT calibration, all students' test results from 1992 to 1996 are used to estimate test statistics. The students' test information used in this study is a subset of this larger dataset.

Assessment System Corporation's (1995b) ITEMAN software package is used for classical test analysis. Table 10 presents a summary of the analyses for each test. Table 10a shows results for the listening test. For this analysis, both forms A & B are combined. Table 10b presents reading test analyses. Because reading subtests all have linking items, it is not necessary to combine forms; thus classical test analyses are conducted on each version and form of the reading test. Appendix A displays results of both classical and IRT items analyses for all exams.

Table 10a: Summary of Classical Test Analysis for Listening Subtests

	Listening Comprehension Test Versions	
Statistics	L92-1	L92-2
Number of items	50	53
Number of students	897	813
Mean	33.67	28.73
Median	35.00	28.00
Standard Deviation	7.62	7.78
Minimum score	10	13
Maximum score	50	50
Mean P-value	.673	.542
Mean P-Biserial Correlation	.493	.419
Cronbach Alpha	.855	.830
SEM	2.90	3.21

Table 10b: Summary of Classical Test Analysis for Reading Subtests

	Reading Comprehension Test Versions and Forms							
Statistics	91-1a	91-1b	91-2a	91-2b	91-3a	91-3b	92-1a	92-1b
Number of items	51	---	49	---	53	---	50	---
Number of students	331	141	320	295	197	174	189	190
Mean	27.25	34.38	34.56	34.65	34.80	34.40	33.01	31.78
Median	29.00	36.00	36.00	36.00	35.00	35.00	35.00	34.00
Standard Deviation	12.72	8.83	7.73	7.81	8.32	8.18	9.48	9.85
Minimum score	8	13	7	8	13	14	5	10
Maximum score	50	48	49	48	51	52	47	49
Mean P-value	.535	.674	.705	.707	.657	.649	.660	.636
Mean P-Biserial Correlation	.660	.545	.533	.543	.490	.493	.582	.584
Cronbach Alpha	.945	.891	.866	.870	.867	.861	.906	.910
SEM	2.97	2.91	2.83	2.82	3.04	3.06	2.90	2.95

All tests have between 49-53 items. The test with the lowest average P-value is R91-1a (.535), and R91-2b has the highest value (.707). Generally, average P-values are moderately high but well within expected ranges. Mean point biserials provide a correlation between the average item and test performance. The higher the biserial for a particular item, the better that item discriminates between those who perform well on the total test and those who do not. The magnitude of the biserial, however, is conditioned upon an item's P-value (Allen and Yen, 1979). Mean point biserials range between .490 (R91-3a) and .660 (R91-1a). This range is satisfactory.

All tests but L92-2 have median values greater than their means. Having greater medians than means is indicative of negatively skewed distributions. Negatively skewed distributions are not unexpected. Students participating in these tests have generally high levels of English proficiency; additionally, these exams are used as end of course tests.

On average exams have 51 items and scoring ranges (difference between minimum and maximum score) between 36 to 43. With similar numbers of items and score ranges, standard deviations should not greatly differ across forms. With one exception, R91-1a, they do not. Standard deviations range between 7.62 (L92-1) to 12.72 (R91-1a).

R91-1a has such a large standard deviation because it has a bimodal distribution. R91-1a's bimodality is a function of how this examination is used. Care is taken to ensure that ELCPT examinations are administered in a rotated fashion. However, it is often the case that international students arrive after formal ELC testing is concluded--often well into the academic semester. Late arrivals need to be evaluated quickly. Instead of taking a 3 1/2 hour exam battery, students are asked to take a shortened battery consisting of an oral interview, a writing test, and a reading test. The reading test used for this purpose is R91-1a. Generally students arriving late to the ELC are either very proficient in English or very limited, hence the bimodal distribution.

Reliability coefficients (Cronbach Alpha) for listening and reading tests range from .830 (L92-2) to .945 (R91-1a). The obtained reliability estimates are high and very acceptable. Standard errors for exams ranged between 2.82 and 3.21, again acceptable estimates. Overall, the exams used for this study exhibit very positive classical test characteristics. These positive classical test characteristics provide reasonable evidence to move on to the next stage of analysis--conducting a one-parameter item response theory calibration.

4.2 Item Response Theory Calibration

One benefit of item response theory (IRT) models is the creation of test-free person ability estimates (Hambleton, Swaminathan and Rogers, 1991). The focus of the research reported here is to identify how tests track student growth. Having estimates that identify underlying traits greatly aids in this investigation. However, IRT is reliant upon several key assumptions: unidimensionality, local independence, and non-speeded test conditions (Lord, 1980, Hambleton and Swaminathan, 1985, and Hambleton, et al., 1991). To check unidimensionality, a factor analysis is conducted using each test's tetrachoric correlation matrix. Scree plots are then examined to see if major factors are present, which are indicative of unidimensional constructs. A discussion of ELC test design is provided to explain how non-speeded test conditions are met.

Tetrachoric correlations are computed for all tests used in this study. Tetrachoric correlation matrices are then used to conduct exploratory factor analyses on each test to establish whether unidimensionality is present. Tetrachoric correlation matrices are used since phi coefficient matrices have been shown to cause spurious results when using factor analysis (Crocker and Algina, 1986; Hambleton and Swaminathan, 1989; Hambleton, et al. 1991). Reckase (1979) and more recently Stout (1990) and Nandakumar (1991) argue that it is not necessary to hold to an extremely conservative view of variance needed to claim unidimensionality. Recakase holds that the first factor need only be 20% of the total variance. However, it should be clear that to meet IRT assumptions a dominant factor must be clearly present. The 20% variance value is used here for IRT assumptions. Table 11 presents factor analysis results for each test. Columns in Table 11 display total variance for each test, factor loadings for each first factor, and proportion of variance explained by first factors. Appendix B presents scree plots and factor loadings greater than one for each analysis.

Table 11: Factor Analytic Results for Tests Used in Study

Test	Total Variance	1^{st} Factor Loading	Proportion Variance
L92-1	50	11.82	23.65%
L92-2	53	10.10	19.06%
R91-1a	51	23.56	46.19%
R91-1b	51	14.80	29.02%
R91-2a	49	13.56	27.63%
R91-2b	49	13.80	28.17%
R91-3a	53	12.57	23.72%
R91-3b	53	12.88	24.30%
R92-1a	50	16.39	32.78%
R92-1b	50	16.74	33.48%

All tests exhibit first factor variances greater than 20% with the exception of L92-2, which has 19.06%. Glancing at L92-2's scree plot (Appendix B) reveals that this exam may have other significant factors, which implies multidimensionality. Buck (1991) writes that listening tests may inherently have at least two distinct dimensions: linguistic ability to decode a listening passage and the propositional knowledge to interpret it. The proportion of variance displayed by L92-1--the other listening test in this sample--while not below 20%, has the next lowest first factor of all tests. This may lend credence to Buck's assertions. It is important to note that the first factor eigenvalue for L92-2 is double that of its second (10.10 versus 5.06). Clearly there is a dominant factor in L92-2's analysis. While L92-2's first factor variance does not meet Reckase's criteria, it is close. Further, listening tests may have unique multidimensional characteristics. For this study, we will assume that L92-2 meets an "essential dimensionality" requirement for using IRT. All other exams meet Reckase's criteria. Evidence from factor analyses presented here suggests that these examinations have dominant main factors and that IRT analysis is appropriate.

To ensure examinees have sufficient time, all ELC examinations are piloted. ELC pilot tests are conducted on students from a wide range of English language abilities: beginning to advanced. It is expected that multiple-choice items will take approximately one minute to complete per item. This time limit is evaluated during piloting. Typically, the time required for 80% of the examinees to complete the test is used as benchmark for final timing. An additional 5 minutes is added to ensure students are not rushed. Hambleton, et al. (1991) write that if 75% of examinees complete a test and if 80% of test items are completed by all examinees, speed is assumed to be inconsequential (p.57). Thus, these tests are assumed to meet IRT's non-speeded test conditions.

Since ELCPT listening and reading comprehension tests appear to meet the unidimensionality and non-speeded test condition assumptions, a Rasch model IRT calibration is conducted. Item difficulty estimates and fit statistics are displayed in Appendix A.

4.3 Equating Tests and Person Ability Estimates

As part of the IRT calibration process, RASCAL allows users to specify the types of output desired. RASCAL provides item difficulty estimations, fit statistics, item mapping information, and person ability estimates. Students used in this study have unique person ability estimates for each term an exam is taken. However, it is unlikely that examinees have taken the same test twice. Therefore, there is a need to place all person ability estimates upon a common scale. The process of placing ability estimates or item difficulties upon a common scale is called test equating (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991; Wright and Stone, 1979). The equating procedure used here is a variant of that mentioned in chapter 5 of Wright and Stone (1979).

All ELC reading tests are linked through ten common vocabulary items. Conducting a common item equating procedure for reading tests would place all items--and ability estimates--upon a common scale. However, listening tests do not share common items. It is necessary to link listening tests through a common person equating procedure.

Table 12: Reading Test Linking Estimates and Linking Constants

Linking	Tests							
Items	R91-1b	R91-1a	R91-2a	R91-2b	R91-3a	R91-3b	R92-1a	R92-1b
1	1.10	-0.23	1.46	1.20	0.92	1.17	1.30	1.11
2	1.06	0.39	0.66	0.70	0.73	0.46	0.97	0.95
3	-0.55	-0.80	-0.87	-0.55	-0.42	-0.62	-0.38	-0.40
4	0.21	0.22	-0.49	0.26	-0.07	-0.18	-0.13	-0.09
5	-0.65	-0.25	-0.92	-1.13	-1.14	-0.95	-1.26	-1.30
6	-0.60	-0.44	-1.42	-1.54	-1.19	-1.23	-1.42	-1.30
7	1.13	0.42	0.70	1.17	1.09	0.99	0.76	1.14
8	1.77	-0.11	1.59	1.69	1.53	1.26	1.68	1.33
9	-0.02	-0.11	-0.21	-0.66	-0.04	-0.27	-0.42	-0.30
10	0.65	0.25	0.84	0.49	0.97	0.91	0.91	0.74
Avg	0.41	-0.07	0.13	0.16	0.24	0.15	0.20	0.19
Linking Constant		0.48	0.28	0.25	0.17	0.26	0.21	0.22

Table 12 displays all 10 linking item estimates for reading tests. The average linking item logit estimate per test and linking constant for each exam is also displayed. R91-1b is the anchor test used to link all reading tests together. The equation below presents the linking constant formula.

$$\text{LinkingConstant} = \frac{\sum\left(Logit_{Anchor} - Logit_{t\arg et}\right)}{\text{Numberof LinkingItems}} \tag{1}$$

For equation 1, $Logit_{Anchor}$ represents the item difficulty estimate (in logits) for anchor items (R91-1b) and $Logit_{target}$ represents the difficulty estimate of the same linking items for the test to be equated (target).

Linking constants are then used to create a common scale for examinee scores. For example, if a student received an overall ability estimate (θ) of 1.00 on R92-1a, their adjusted ability estimate (θ_{adjusted}) would be 1.21 (1.00 + 0.21). All students' reading test ability estimates were placed upon a common scale using this procedure.

Listening tests do not have common items; therefore, a common item equating procedure could not be used. To link tests together a common person equating model is needed. This procedure is similar except student ability estimates are used instead of item difficulty estimates. To conduct a common person equating procedure, students who took both L92-1 and L92-2 at the same point in time had to be identified. In spring of 1994, 36 students were found to fit that criterion. Students in this linking sample were visiting European students given a readministration at the request of their chaperone. Table 13 shows ability estimates on each listening test for participating students, as well as displaying the linking constant.

Table 13: Listening Test Linking Estimates and Linking Constant

Student	L92-1	L92-2	Difference	Student	L92-1	L92-2	Difference
1	1.80	2.07	0.27	19	2.29	1.70	-0.59
2	1.02	-0.26	-1.28	20	1.58	1.37	-0.21
3	1.20	0.79	-0.41	21	1.58	1.37	-0.21
4	1.80	1.37	-0.42	22	1.38	0.79	-0.59
5	1.79	0.26	-1.53	23	1.38	1.07	-0.31
6	2.03	1.37	-0.66	24	1.58	1.37	-0.21
7	1.58	1.07	-0.51	25	2.03	1.37	-0.66
8	0.68	2.07	1.38	26	1.20	0.00	-1.20
9	2.29	1.37	-0.92	27	2.03	2.50	0.47
10	1.20	1.37	0.18	28	2.03	2.50	0.47
11	1.38	1.70	0.32	29	3.45	1.70	-1.75
12	1.20	0.52	-0.68	30	1.58	1.37	-0.21
13	0.52	0.26	-0.26	31	1.58	1.37	-0.21
14	2.59	0.79	-1.80	32	0.05	1.07	1.02
15	1.38	1.70	0.32	33	2.59	1.70	-0.89
16	2.59	2.07	-0.53	34	1.38	0.26	-1.13
17	2.29	2.50	0.20	35	2.59	2.07	-0.53
18	0.52	0.26	-0.26	36	3.45	3.04	-0.40
				Average	1.71	1.33	
				Linking Constant			-0.38

Equation 2 presents the common person equating linking constant formula.

$$\text{Linking Constant} = \frac{\sum\left(\text{Ability}_{L922} - \text{Ability}_{L921}\right)}{\text{Number of Persons}} \qquad (2)$$

For this linking procedure, L92-2 was used as the anchor examination. If a student received a $\theta = 1.00$ on L92-1, the adjusted estimate would be 1.00 -.38 or $\theta_{\text{adjusted}} = 0.62$. All student scores were placed upon a common scale using this procedure.

Once ability estimates were obtained, student demographic information and term variables were added. The next section presents statistical summaries of this combined information.

4.4 Descriptive Summaries of Ability Estimate Data

Student ability estimates, demographic information and term variables are combined into a common file for analysis using SPSS 7.5.1. Ability estimates used for analyses are the logistic probability estimates (logits) obtained from the RASCAL program. However, a note of caution should be offered. The listening and reading tests are calibrated separately; thus, logits cannot be directly compared between listening and reading tests. They are not on the same scale.

Recall that the demographic variables are sex, age, language group and academic status. The variable "term" represents the academic term (more specifically, academic semester) an exam score is reported. The following illustrates term coding. An examinee arrives in fall of 1994 and takes an initial placement test. At the end of the fall term, she takes a parallel exam. The examinee has one more recorded score--summer of 1995. This examinee's term coding would be 0, 1, 2, 3. The examinee's ability estimates would be recorded for terms 0, 1, and 3. Term 2 would have a missing score. For this study the term value of 0 indicates initial score. Table 14 presents the frequency and average logits per term.

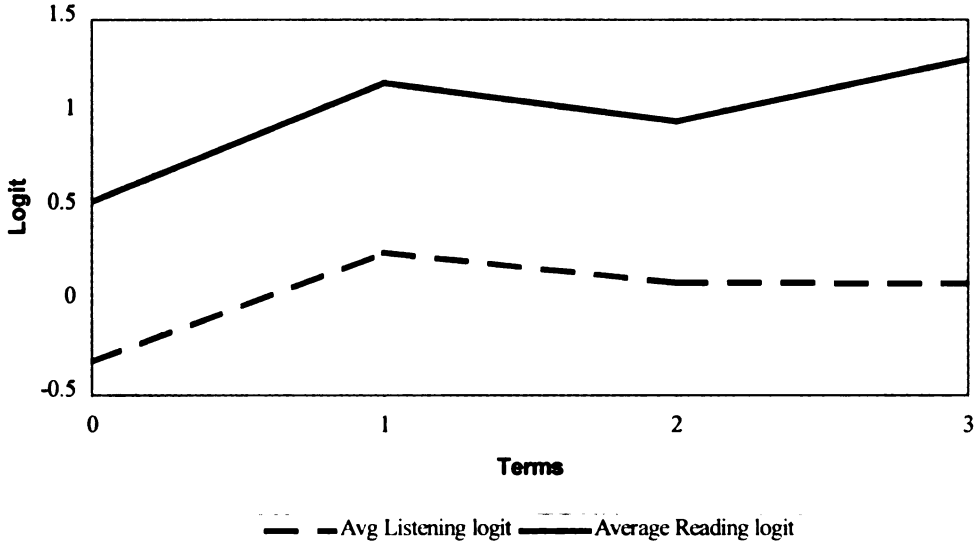
Table 14: Frequency and Mean Logit per Term

Term	Listening Test		Reading Test	
	Mean Logit	N	Mean Logit	N
0	-.325	199	.532	213
1	.260	152	1.162	208
2	.098	24	.959	60
3	.096	32	1.288	15
4	.031	9	.849	1
5	.275	13	---	---
6	-.550	6	---	---
7	.358	1	---	---
8	-.380	1	---	---
Total	-.044	437	.871	497

Notice for the listening test that one examinee has 8 full terms recorded (practically 3 years). The reading test has a maximum 4 terms (1 1/2 years). However, most scores for both listening and reading tests are within the first three terms (listening, 93.2% and reading 99.8%). Since a majority of student scores occur within three terms, further descriptive analyses only present terms 0 through 3.

Figure 1 shows average logit for the first three terms on both tests.

Figure 1: Average Logit per Term--Listening and Reading Tests



Both tests' average logit scores increase between the initial and first term and decrease between the first and second term. Between the second and third term average listening scores go down slightly, and reading scores increase. Both tests exhibit non-linear-like growth trajectories.

Since a large majority of students have only two time points (typically meaning they have completed ELC course requirements), there may be differences in growth trajectories for students who complete in one term and those who take more than one term. Table 15 displays the mean subtest logit for students with two and more than two time-points.

Table 15: Mean Subtest Logit for Students with Two and More than Two Time-points

Term	Statistic	Listening Test		Reading Test	
		1 Term	>1 Term	1 Term	>1 Term
0	Mean	-.170	-.602	.721	-.068
	N	128	71	162	51
1	Mean	.378	-.368	1.375	.370
	N	128	24	164	44
2	Mean		.098		.959
	N		24		60
3	Mean		.096		1.288
	N		32		15
4	Mean		.031		.849
	N		9		1
5	Mean		.275		
	N		13		
6	Mean		-.550		
	N		6		
7	Mean		.358		
	N		1		
8	Mean		-.380		
	N		1		
Total	Mean	.104	-.252	1.050	.529
	N	256	181	326	171

There are dramatic differences between students who finish in one term and those who take longer. On both the listening and reading comprehension tests, students who complete in one term have much higher average initial logit values. Further, those who complete in one term have much steeper growth rates. The differences in trajectories can be seen in Figure 2 which compares the average growth trajectories of students completing in one term and those taking more than one term.

Figure 2a: Listening Comprehension Test Growth Trajectory Comparing Students Enrolled for One Term and More than One Term

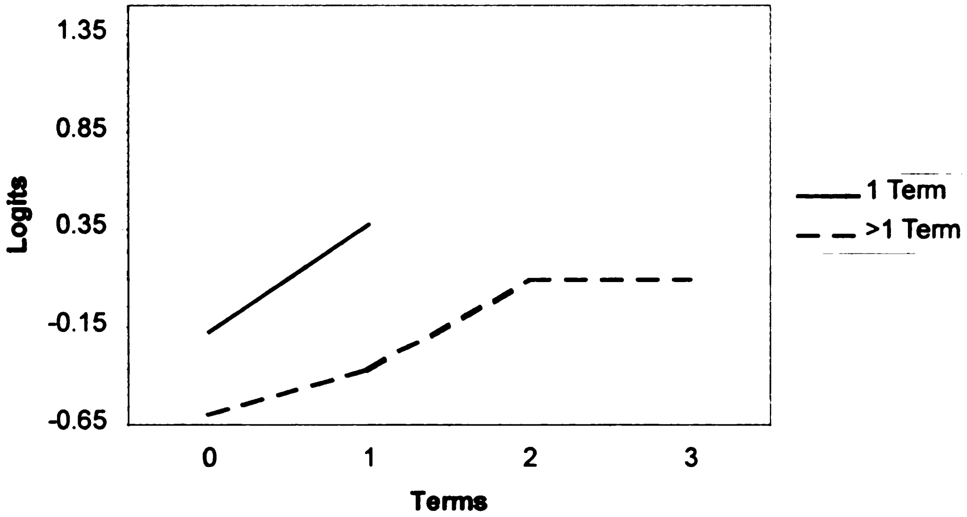
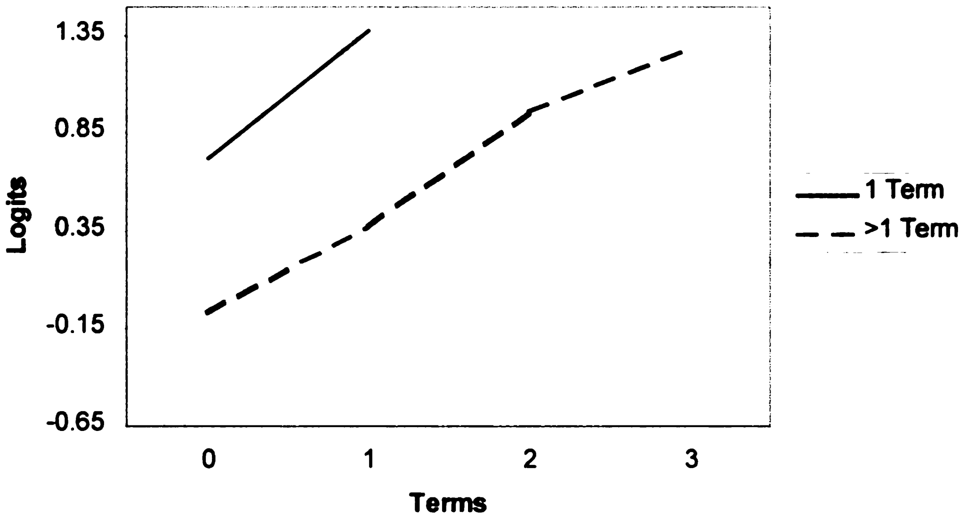


Figure 2b: Reading Comprehension Test Growth Trajectory Comparing Students Enrolled for One Term and More than One Term



For display purposes, the listening and reading comprehension test graphs in Figure 2 have the same range of logit values on the y-axis. It is important to note that these tests cannot be directly compared to each other since they are not on the same IRT logit scale. However, trends can be compared, and in both listening and reading tests, students who finish in one term have higher starting average logits and steeper growth curves. This seems to justify the use of the level two variable "More than two terms."

Table 16 portrays the differences in logit scores by sex for each exam.

Table 16: Average Subtest Mean Logit by Term by Sex

Term	Listening Test		Reading Test	
	Male	Female	Male	Female
0	-.350	-.289	.475	.620
1	.257	.265	1.141	1.212
2	.019	.165	.788	1.142
3	.024	.202	1.255	1.325

Differences in average logit between sexes are slight. Two noticeable exceptions are females tend to score higher across terms and females display less of a drop in scores between the first and second terms on both tests.

Examinees' ages range between 17 and 48, with an average age of 23.87 for listening and 24.43 for reading (see Table 4). While age is a continuous variable, it is interesting to parcel age into distinct categories to investigate trends. Table 17 displays average logit per age group. Four groups are created: those younger than 21, those between 21 and 25, those between 26 and 30, and those older than 30.

Table 17a: Listening Test--Mean Logit by Age Group

Age Group	Statistic	Term				Total
		0	1	2	3	
<21	Mean	-.248	.424	.086	-.171	.037
	N	37	29	8	5	79
21-25	Mean	-.330	.291	.060	.140	-.033
	N	110	90	11	15	226
26-30	Mean	-.355	.094	.339	.406	-.088
	N	40	25	4	9	78
>30	Mean	-.407	-.152	-.344	-.609	-.345
	N	12	8	1	3	24
Total	Mean	-.325	.260	.098	.096	-.048
	N	199	152	24	32	407

Table 17b: Reading Test--Mean Logit by Age Group

Age Group	Statistic	Term				Total
		0	1	2	3	
<21	Mean	.356	.970	.757	.831	.683
	N	39	38	17	4	98
21-25	Mean	.523	1.201	1.021	1.178	.883
	N	112	108	26	8	254
26-30	Mean	.626	1.099	1.031	.156	.880
	N	41	42	13	1	97
>30	Mean	.720	1.452	1.186	3.208	1.177
	N	21	20	4	2	47
Total	Mean	.532	1.162	.959	1.288	.871
	N	213	208	60	15	496

A very interesting trend is that examinees over 30 perform poorest on the listening test, i.e., they make the smallest overall gain, yet perform best on the reading test. In a similar fashion, the youngest group of students (<21) perform best on the listening test and poorest on the reading test. This trend certainly bears further investigation.

Table 18, below, displays average logit per term by language group. Several interesting trends can be seen from this table. One prominent finding is language groups display a diverse array of trajectories across terms.

Table 18a: Listening Test--Mean Logit by Term by Language Group

Language Group	Statistics	Term				
		0	1	2	3	Total
Other Language	Mean	-.029	.598	.360	.178	.247
Groups	N	25	20	2	5	52
Middle Eastern	Mean	-.467	.418	-.085	-.113	-.121
	N	19	12	2	5	38
Chinese	Mean	-.362	.095	---	.690	-.078
	N	23	20	---	4	47
Japanese	Mean	-.354	.179	.155	-.131	-.106
	N	42	32	6	8	88
Korean	Mean	-.353	.220	.063	.104	-.082
	N	90	68	14	10	182
Total	Mean	-.324	.260	.098	.096	-.048
	N	199	152	24	32	407

Table 18b: Reading Test--Mean Logit by Term by Language Group

Language Group	Statistics	Term				
		0	1	2	3	Total
Other Language	Mean	.773	1.598	1.209	1.540	1.196
Groups	N	34	34	3	2	73
Middle Eastern	Mean	-.023	.432	1.020	1.744	.348
	N	16	14	5	1	36
Chinese	Mean	.695	1.318	.948		.993
	N	31	29	5		65
Japanese	Mean	.293	.870	.563	.346	.562
	N	45	42	16	4	107
Korean	Mean	.605	1.198	1.131	1.639	.965
	N	87	89	31	8	215
Total	Mean	.532	1.162	.959	1.288	.871
	N	213	208	60	15	496

For the listening test, the "Other language groups" group has the highest initial listening test score (initial logit=-0.029). The lowest average initial listening score and the lowest initial test to first term gain score is from the Chinese language group.

Japanese and Korean language groups have similarly low initial listening scores. The Middle Eastern students make the largest initial to first term gain (gain=0.885 logits).

On the reading test (Table 18b), the "Other language groups" group has the highest average initial reading test score and the highest initial to first term gain. Conversely, the Middle Eastern students have the lowest initial average score and the lowest initial to first term gain.

Like language groups, average logit per term fluctuates greatly across students with different academic statuses (see Table 19). On both reading and listening tests, students studying English have lower average scores. That is to be expected, since students are coming to study English and presumably have low proficiency. Interestingly, graduate students do much better on reading tests than on listening tests. Another interesting trend is graduate students do not exhibit a first to second term slump (loss in average logit) on the reading test.

Table 19a: Listening Test--Mean Logit by Term by Academic Group

Academic Status	Statistic	Term				Total
		0	1	2	3	
Studying	Mean	-.380	.112	-.021	-.068	-.152
English	N	97	71	14	17	199
Undergraduate	Mean	-.233	.478	.148	.516	.121
Student	N	64	52	8	10	134
Graduate	Mean	-.336	.233	.734	-.183	-.074
Student	N	38	29	2	5	74
Total	Mean	-.324	.260	.098	.096	-.048
	N	199	152	24	32	407

Table 19b: Reading Test--Mean Logit by Term by Academic Group

Academic Status	Statistic	Term				Total
		0	1	2	3	
Studying English	Mean	.339	.983	.808	.858	.689
	N	80	80	29	6	195
Undergraduate	Mean	.608	1.247	1.006	1.525	.958
	N	82	80	22	8	192
Graduate	Mean	.685	1.276	1.332	1.971	1.013
	N	47	44	9	1	101
Missing	Mean	1.048	1.809			1.428
	N	4	4			8
Total	Mean	.532	1.162	.959	1.288	.871
	N	213	208	60	15	496

To review, the average gain per term for both listening and reading tests exhibits non-linear characteristics (see Figure 1) when all students are grouped. However, when students who have more than two time points are compared with students who have only two times points, we see a different picture (see Table 15 and Figure 2). Students with more than two time points on the listening test show clear non-linear characteristics, but the same class of students on the reading test exhibits more linear trajectories.

On average female subjects score slightly higher than male subjects. Examinees over thirty score best on the reading test and poorest on the listening test, while the youngest test-takers perform best on the listening test and poorest on the reading test. Language groups greatly differ in their scores and score trajectories. On both tests, students from other language groups have higher initial scores. Middle Eastern students exhibit the highest initial to first term gain on the listening test but the lowest initial to first term gain on the reading test. Test performance based on academic status also varies greatly. Students attending the ELC to study English tend to have lower scores than graduate or undergraduate students. Graduate students perform much better on reading tests than they do on listening tests.

CHAPTER 5: HIERARCHICAL LINEAR MODELING ANALYSIS OF STUDENT PERFORMANCE

5.1 Background to Hierarchical Linear Models

Chapter 4 describes student performance over time by demographic group. Several differences were noted: change over time and differences between sexes, age groups, language groups and academic status. Once differences are discovered, a natural next step is to conduct inferential analyses to determine if observed differences are in fact significant. However, data used in this study are hierarchical. That is, students' ability estimates are recorded over time (terms), and students have unique demographic variability that might affect performance at each time-point. Neither traditional regression analyses nor multivariate repeated measurement techniques (e.g., repeated measures analysis of variance) adequately specify this type of hierarchically based model. These techniques can determine significant differences between time points or determine if demographic variables affect student performance within time points. But they cannot adequately specify differences within and across times points taking all available information into account. Bryk and Raudenbush, (1992) argue that hierarchical linear modeling approaches are superior to traditional repeated measurement techniques because they allow growth to be specified a priori, have flexible data requirements, allow for flexible specification of covariance structures, and can be directly related to multivariate repeated measurement models (p. 133). Due to this study's data structure, a hierarchical linear modeling (HLM) approach seems most appropriate. What follows is a brief description of HLM as it relates specifically to studying change over time.

Growth models using HLM can be conceived as two sets of hierarchically related regression equations. The first equation represents the repeated-observations level or simply level one. This level can be expressed by the follows:

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\, a_{ti} + \Pi_{2i}\, a^{2}_{ti} + e_{ti}. \tag{3}$$

For our purposes, Y_{ti} represents the observed logit value for examinee *i* at time *t*; Π_{0i} is the y-intercept, or value of Y_{ti} when *a* is zero; Π_{1i} is the observed linear growth trajectory parameter; Π_{2i} is the quadratic growth trajectory parameter, and e_{ti} is the random level 1 error which is assumed to have a mean of zero and a variance of σ^2. For growth models, the unit-time measure *a* can represent any meaningful unit of time: hour, day, month, year, etc. For this study, *a* represents academic term. An interesting feature of the level one model is the degree polynomial associated with *a.* Adding polynomials (e.g., the quadratic term $\Pi_{2i}\, a_{ti}^{2}$) allows researchers to investigate non-linear models--a useful feature since many growth analyses exhibit non-linear characteristics (e.g., Huttenlocher, et. al., 1991; Pearson, et. al., 1993; Larsen-Freeman, 1993).

The next level (level two) is associated with person variables. The model at this level can be expressed as

$$\Pi_{0i} = B_{00} + B_{0q}\, X_{qi} + r_{0i}. \tag{4a}$$

$$\Pi_{1i} = B_{10} + B_{1q}\, X_{qi} + r_{1i}, \tag{4b}$$

$$\Pi_{2i} = B_{20} + B_{2q}\, X_{qi} + r_{2i}. \tag{4c}$$

Notice that the parameter estimates represented in level one now become outcome variables in level two. For level 2, X_{qi} represents person characteristics (e.g., sex or age). B_{00} is the average y-intercept across all persons; B_{10} represents average instantaneous linear growth trajectory across all persons, and B_{20} is the average change in acceleration of each growth trajectory; B_{0q}, B_{1q}, and B_{2q} represent the effects of X_{qi} on each growth parameter. Finally, r_{0i}, r_{1i}, and r_{2i} are random effects, which are assumed to have means of zero and are multivariate normally distributed with variance-covariance matrix *T*. For a much more detailed description of HLM growth models see Bryk and Raudenbush (1992) Chapter 6.

The goal in conducting HLM analysis is to identify the most parsimonious model given the observed data. The HLM2 computer program (Bryk and Raudenbush, 1988) is used for HLM analyses reported here. A five-step procedure is followed to identify the best HLM model:

Step 1: Specify possible models at level 1 and 2,

Step 2: Evaluate fixed parameter estimates (e.g., B_{00} or B_{10}),

Step 3: Evaluate random parameter estimates (e.g., r_0 or r_1),

Step 4: Examine the correlation between initial status and change, and

Step 5: Identify variance explained by each new model.

Step one introduces a model to be tested. In step two, initial status and change parameter estimates of the specified model (fixed effects) are investigated. In HLM2 fixed effects estimates are evaluated using t-tests. At step three, we evaluate random parameter estimates using chi-square statistics. Significant random effects reflect substantial variation not explained by the examined model; thus, further exploration of other HLM models may be warranted. The HLM2 program provides initial status and rate of change correlation coefficients (Step 4) as part of the program's output. In step five variances are calculated using the following equation (found in Bryk and Raudenbush, 1992, p.74).

These variances provide evidence of the amount of variation explained by a proposed model.

$$\text{Variance Explained} = \frac{\hat{\tau}_{qq}(\text{random regression}) - \hat{\tau}_{qq}\ (\text{fitted model})}{\hat{\tau}_{qq}(\text{random regression})} \tag{5}$$

5.2 Modeling Listening Growth Trajectory

The listening test growth model is evaluated first. The initial model for the HLM listening test analysis is shown below:

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\ (\text{linear term})_i + e_{ti}\,, \tag{6a}$$

$$\Pi_{0i} = B_{00} + r_{0i}\,, \tag{6b}$$

$$\Pi_{1i} = B_{10} + r_{1i}\,. \tag{6c}$$

The linear individual growth model (equation 6) is the simplest model for evaluating change over time. Specifying the model is the first step in the HLM analysis. Table 20 presents the results of the listening test linear individual growth model analysis.

Table 20: Listening Test--Linear Model

```
Final estimation of fixed effects:
----------------------------------------------------------------------------
   Fixed Effect      Coefficient   Standard Error  T-ratio    P-value
----------------------------------------------------------------------------
For    INTRCPT1, Π0
   INTRCPT2, B00       0.101708        0.050709        2.006     0.045
For    LINEAR slope, Π1
   INTRCPT2, B10       0.268724        0.025613       10.492     0.000

Final estimation of variance components:
----------------------------------------------------------------------------
Random Effect        Standard     Variance    df    Chi-square    P-value
                     Deviation    Component
----------------------------------------------------------------------------
INTRCPT1,        R0   0.60254      0.36305   197     491.12817    0.000
  LINEAR slope,  R1   0.13400      0.01796   197     237.85418    0.025
 level-1,        E    0.48642      0.23660

Tau (as correlations)
 INTRCPT1  1.000  0.984
   LINEAR  0.984  1.000
```

The next step in HLM growth model analysis is to determine which fixed effect parameter estimates are significant. From Table 20, the estimated mean intercept, B00, and mean linear growth rate, B10, are 0.102 and 0.269, respectively. The intercept (B00) and slope (B10) coefficients have high t-ratios and are significant. This finding indicates that these fixed-effect parameters should be retained in this model.

It is important to note that the intercept term, B00, used in these models, is centered in SPSS prior to generating HLM files. Centering reduces the possibility of collinearity when combining linear and quadratic terms. For listening test data, the average term value is 1.3. Recall that term values are represented as 0, 1, 2, 3, etc. To center terms, subtract the average term estimate (1.3) from each term value. The centered initial status or starting test value is

-1.3 (0 - 1.3). The first term centered value is -.3 (1 - 1.3) and so on. Using fixed-effect parameter estimates and centered term values, predicted logits per term can be calculated. For initial status, the estimated value is 0.102 + 0.269 (-1.3), which equals -0.248. The estimated value for first term logit is 0. 102 + 0.269 (-.3), which equals 0.071. Remaining estimates for later terms are calculated in similar fashion.

The third step in HLM analysis involves the evaluation of a model's random effects. The estimated random effect variances for the linear growth model are 0.363 for R0 and 0.018 for R1. Chi-square statistics for the intercept variance component (R0) is 491.128 which is significant at the p <.000 level. The linear slope Chi-square estimate (R1) is 237.854 and has a p-value of 0.025. The null hypotheses for these analyses are the variance of R0 = 0 and the variance of R1 = 0. In both cases, the null hypotheses are rejected. This means that there is substantial variability in the intercept (initial status) and slope (linear growth rate) portions of this model. These findings provide evidence for estimating a more complex linear model of listening growth in both the intercept and slope parameters.

Next, the correlation between the initial status and linear growth parameter is examined. The correlation between intercept and slope parameter is .984. This means that students with high initial listening test scores also have high linear growth trajectories. Similarly, students with low initial listening test scores have low linear growth trajectories. The last step is to calculate the variance explained by each new model. Since the linear growth model is the starting model from which all other more complex linear models are evaluated, no comparison is provided. The random effect variance components for this model will be used to measure changes in all other linear models.

Prior to adding variables to the linear model, another basic model may also fit the data--a curvilinear model (as suggested by Larsen-Freeman, 1993 and Long, 1990). This curvilinear model is expressed as follows:

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\text{ (linear term)}_i + \Pi_{2i}\text{ (quadratic term)}_i + e_{ti}, \qquad (7a)$$

$$\Pi_{0i} = B_{00} + r_{0i}, \qquad (7b)$$

$$\Pi_{1i} = B_{10} + r_{1i}, \qquad (7c)$$

$$\Pi_{2i} = B_{20} + r_{2i}. \qquad (7d)$$

In Equation 7, one more component is added to the linear model--the quadratic term. The quadratic term is created by squaring the linear term. This new element in the equation adds another fixed effect (Π_{20}) and random effect (r_{2i}) into the HLM analysis. The quadratic growth model for listening has three main fixed effects: initial status or starting listening logit (B_{00}), instantaneous growth rate (B_{10}), and acceleration of growth rate (B_{20}). Table 21 presents the results of this analysis.

Table 21: Listening Test--Quadratic Growth Model

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio    P-value
 ----------------------------------------------------------------------
 For   INTRCPT1, Π0
    INTRCPT2, B00        0.228333        0.061810         3.694     0.000
 For   LINEAR slope, Π1
    INTRCPT2, B10        0.233847        0.042283         5.531     0.000
 For     QUAD slope, Π2
    INTRCPT2, B20       -0.133118        0.036351        -3.662     0.000

 Final estimation of variance components:
 -----------------------------------------------------------------------------
 Random Effect         Standard    Variance      df    Chi-square    P-value
                       Deviation   Component
 -----------------------------------------------------------------------------
 INTRCPT1,          R0  0.81667        0.66695    37  114984.79263    0.000
   LINEAR slope,    R1  0.33495        0.11219    37   39445.02258    0.000
     QUAD slope,    R2  0.41338        0.17088    37  108259.64377    0.000
  level-1,           E  0.01501        0.00023

Note: The chi-square statistics reported above are based on only 38 of
199 units that had sufficient data for computation.

Tau (as correlations)
 INTRCPT1  1.000  0.082 -0.618
   LINEAR  0.082  1.000  0.306
     QUAD -0.618  0.306  1.000
```

All fixed effects in Table 21 (B00, B10 and B20) have high t-ratios, indicating that the quadratic growth model also seems to account for variance in the listening data. The quadratic model, like the linear model, may also be an appropriate method of modeling listening growth. The main difference between the linear model and the quadratic model is the acceleration rate parameter (B20). Note that the obtained estimate for B20 is negative (-0.133). This means that on average students' listening growth rates do not continually increase but taper off.

Similarly, all random effects that are tested have high chi-square statistics and p-values less than p<.01. However, a very important note of caution is in order. To calculate random effects hypothesis tests, only 38 students' listening score trajectories could be used. A substantial number of students are not used in generating hypothesis tests (over 80%). The main reason for the loss of data is most students have only two time points. It is not possible to model quadratic effects with less than three time points. Essentially, only those students who have more than two time points are used to generate estimates and tests for this quadratic model. Theoretically, the quadratic model is more appealing especially given Larsen-Freeman's (1993) and Long's (1990) comments about language acquisition being non-linear. The students remaining in the sample (n=38) do seem to exhibit quadratic tendencies, but further exploration of this and other more complex models would be limited to only those staying at the ELC for more than one term. Several quadratic models were attempted (but not reported here); however, none had substantial findings. This is primarily due to the small number of students and limited data points. Because of limited data and the limited inferences made with only 38 students, pursuit of a quadratic model is abandoned.

Once the level one model is selected, in this case the linear model, the next step is to add level two variables. The first model explored adds the variable "more than one term" (More). This variable relates to the number of terms a student stays at the ELC. Figure 2a (Chapter 4) showed two distinct trajectories for those who stay at the ELC: one for those who stay one term and one for those who stay longer than one term. Descriptively and graphically, there seems to be substantial differences, but are they significant? The next analysis investigates this. Equation 8 presents the model being investigated. Table 22 presents the results of this analysis.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i} \text{ (linear term)}_i + e_{ti}\,, \tag{8a}$$

$$\Pi_{0i} = B_{00} + B_{01} \text{ (More)}_i + r_{0i}, \tag{8b}$$

$$\Pi_{1i} = B_{10} + B_{11} \text{ (More)}_i + r_{1i}\,. \tag{8c}$$

Table 22: Listening Test--Linear Growth Model with Person Variable "More"

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For   INTRCPT1, Π0
    INTRCPT2, B00       0.528288        0.073940        7.145     0.000
        MORE, B01      -0.848428        0.101168       -8.386     0.000
 For   LINEAR slope, Π1
    INTRCPT2, B10       0.526241        0.060629        8.680     0.000
        MORE, B11      -0.337794        0.065279       -5.175     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect       Standard     Variance     df    Chi-square   P-value
                     Deviation    Component
 ----------------------------------------------------------------------
 INTRCPT1,         R0  0.49200    0.24206      196     362.53528    0.000
   LINEAR slope,   R1  0.08445    0.00713      196     205.47520    0.307
  level-1,         E   0.47796    0.22845
```

```
******* ITERATION 1181 *******

Tau (as correlations)
 INTRCPT1  1.000  0.950
   LINEAR  0.950  1.000
```

The fixed effects in Table 22 are all significant, which indicates that the variable "More" should be retained in the model. Students who stay at the ELC longer than one term have an average intercept value of 0.848 logits less than their classmates who stay but one term. That is a substantial amount. Similarly, students who stay longer than one term have linear growth rates that are 0.338 logits less per term than their classmates. Thus students staying more than one term at the ELC typically start at lower listening abilities and have less steep slopes when compared to their one-term colleagues.

The variance component for the intercept's random effect (R0) is significant (chi-square=362.535, p<.001), meaning that there is substantial variation in the intercept. This part of the model deserves further investigation. The slope's random effect is not significant (chi-square=203.08, p=0.307) which means that there is not significant variation in the slope parameter after accounting for the number of terms. The correlation between the intercept and slope parameters is 0.950, which is similar to that found in the original linear model (Table 20). Students having high initial starting scores have commensurately high growth rates. The last step in the evaluation process is to compare the amount of variation explained by the model in the intercept and slope parameters. To calculate the amount of added variation explained by the model, we apply Equation 5 to the parameter estimates in Tables 20 and 22. Thus, 33.3% ([0.363 - 0.242] / 0.363) of the variation in the intercept is explained by the addition of the variable "More." For the slope parameter 60.3% ([.018 -.007] / .018) of the variation in the slope parameter estimate is accounted for by addition of the variable "More."

The non-significant finding in the slope's random effect suggests that further differences in slope variation may now not exist. An additional piece of information is added to Table 22. The number of iterations needed to come to an HLM solution for this analysis was 1181. Bryk & Raudenbush suggest that iteration counts are highly diagnostic of the amount of information in the data (1992, p.202): the fewer the iterations, the more information in the model. Likewise, Bryk & Raudenbush state that the smaller the reliability estimate (e.g., <.05), the more likely the estimated variance is closer to zero. The reliability estimate for the linear component of this model is 0.071.

Because of the slope's non-significant random effects, high convergence iterations, and low reliability estimates, fixing the slope parameter variance to zero may be warranted. That is, instead of allowing the slope parameter to vary randomly, the variance will be fixed to have a linear component for those students staying one term at the ELC (B_{10}) and a linear component for those staying more than one term (B_{11}). Equation 9 presents this new model.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti} \qquad (9a)$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + r_{0i} \qquad (9b)$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i \qquad (9c)$$

The results of this model are displayed in Table 23.

Table 23: Listening Test--Linear Growth Model with Person Variable "More" and Slope Variance Fixed to Zero

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For    INTRCPT1, Π0
    INTRCPT2, B00       0.524214        0.073563         7.126     0.000
        MORE, B01      -0.855466        0.098132        -8.718     0.000
 For    LINEAR slope, Π1
    INTRCPT2, B10       0.522891        0.062751         8.333     0.000
        MORE, B11      -0.369362        0.066067        -5.591     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect       Standard      Variance      df     Chi-square    P-value
                     Deviation     Component
 ----------------------------------------------------------------------
 INTRCPT1, R0        0.44151        0.19493    197      529.51719     0.000
  level-1, E         0.49856        0.24857
```

All fixed effects coefficients are significant as in the previous model. Students who stay one term at the ELC have a predicted intercept logit of 0.524. Students staying longer have a predicted intercept logit 0.856 points less than their one term classmates. Notice that there is only one random effect, R0. The random effect coefficient for the slope is removed when slope variance is fixed to zero. Now there is no longer a correlation between the intercept and slope since there is no slope variability. The variance accounted for by the model changes as well. Since the slope's previously random effects are now fixed, there is no slope variance to compare. It is also important to note that the number of iterations required to fit this model is substantially less than the model with the randomly varying slope (1181 for the random slope versus 8 with the fixed slope). The above analysis will be used as the base model to compare further HLM listening growth models.

There is still substantial variation in the intercept portion of the model. Essentially, the intercept represents students' initial listening test scores (logits). The remaining models investigate whether significant differences exist in students' starting scores based upon demographic information (age, sex, academic status, and language group).

The next model, represented by Equation 10, investigates whether age is a factor in determining students' initial scores. The model presented in Equation 10 examines whether age by More interactions exist in the intercept. Table 24 presents the results of this model's analysis.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i} \text{ (linear term)}_i + e_{ti} , \qquad (10a)$$

$$\Pi_{0i} = B_{00} + B_{01} \text{ (More)}_i + B_{02} \text{ (Age)}_i + B_{03} \text{ (Age x More)}_i + r_{0i} , \qquad (10b)$$

$$\Pi_{1i} = B_{10} + B_{11} \text{ (More)}_i . \qquad (10c)$$

Table 24: Listening Test--Linear Growth Model with Age by More Interaction

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For   INTRCPT1, Π0
    INTRCPT2, B00       0.524276        0.073050        7.177     0.000
        MORE, B01      -0.855140        0.097059       -8.810     0.000
         AGE, B02      -0.027831        0.011564       -2.407     0.016
         AXM, B03       0.006116        0.018066        0.339     0.735
 For   LINEAR slope, Π1
    INTRCPT2, B10       0.523956        0.062743        8.351     0.000
        MORE, B11      -0.370450        0.066052       -5.608     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect      Standard     Variance      df     Chi-square    P-value
                    Deviation    Component
 ----------------------------------------------------------------------
 INTRCPT1, R0       0.43070        0.18550    195      508.40452     0.000
  level-1, E        0.49849        0.24849
```

The added variables in this model are age and age by more interaction (AXM). A note of caution is in order. When interaction terms are present, main effects should not be interpreted. Because the interaction term is not significant, a further reduction in fixed effects is warranted. Equation 10b now becomes

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{Age})_i + r_{0i}\,.$$

Table 25 displays an analysis of this model.

Table 25: Listening Test—Linear Growth Model with Age

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00       0.524117        0.072978        7.182      0.000
        MORE, B01      -0.854922        0.096896       -8.823      0.000
         AGE, B02      -0.025322        0.008862       -2.857      0.005
 For   LINEAR slope, Π1
    INTRCPT2, B10       0.523763        0.062752        8.346      0.000
        MORE, B11      -0.370261        0.066059       -5.605      0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect        Standard      Variance      df     Chi-square    P-value
                      Deviation     Component
 ----------------------------------------------------------------------------
 INTRCPT1,   R0       0.42892       0.18397      196     508.45162     0.000
  level-1,   E        0.49858       0.24858
```

The results in Table 25 indicate that the older a student, the lower their initial listening test score (ability). That is, for every year above average age (average listening age =23.87), a student is predicted to have a 0.025 lower initial listening test logit. Note the reduction in variation in the intercept random effect (R0). In Table 23, the intercept variance is 0.195, and in Table 25, the variance is 0.184. Using Equation 5, this reflects a 5.6% reduction in unexplained variation.

The next model retains the age and More variables and investigates sex by More interactions. Equation 11 presents this model, and Table 26 displays the results.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i} \text{ (linear term)}_i + e_{ti} \,, \tag{11a}$$

$$\Pi_{0i} = B_{00} + B_{01} \text{ (More)}_i + B_{02} \text{ (Age)}_i + B_{03} \text{ (sex)}_i +$$

$$B_{04} \text{ (sex x More)}_i + r_{0i} \,, \tag{11b}$$

$$\Pi_{1i} = B_{10} + B_{11} \text{ (More)}_i \,. \tag{11c}$$

Table 26: Listening Test--Linear Growth Model with Age and Sex by More Interaction

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00      0.534497       0.085867         6.225     0.000
        MORE, B01     -0.916457       0.117492        -7.800     0.000
         SEX, B02     -0.023141       0.100750        -0.230     0.818
         SXM, B03      0.160122       0.165042         0.970     0.332
         AGE, B04     -0.024114       0.009155        -2.634     0.009
 For   LINEAR slope, Π1
    INTRCPT2, B10      0.523748       0.062774         8.343     0.000
        MORE, B11     -0.368416       0.066107        -5.573     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect       Standard      Variance     df     Chi-square    P-value
                     Deviation     Component
 ----------------------------------------------------------------------------
 INTRCPT1,  R0       0.43011        0.18499    194     504.15749     0.000
  level-1,  E        0.49875        0.24875
```

In this model, the sex by More interaction is not a significant predictor of the intercept. A subsequent model is tested removing the interaction term. This model is similar to equation 11 except equation 11b now becomes:

$$\Pi_{0i} = B_{00} + B_{01} \text{ (More)}_i + B_{02} \text{ (Age)}_i + B_{03} \text{ (sex)}_i + r_{0i}.$$

In this model (not reported here) the main effect for sex is not significant. Thus, males and females do not significantly differ in their initial listening test performance.

The next model investigates whether students of different academic statuses staying more than one term differ in their initial listening test scores. Equation 12 presents this model.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti} \qquad (12a)$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{Age})_i + B_{03}\,(\text{Ugr})_i + B_{04}\,(\text{Ugr x More})_i + B_{05}\,(\text{Grad})_i + B_{06}\,(\text{Grad x More})_i + r_{0i} \qquad (12b)$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i \qquad (12c)$$

The model expressed in Equation 12 queries whether undergraduate students staying more than one term (UGXM), graduate students staying more than one term (GRXM), or students studying English (intercept) differ in initial listening test ability.

Table 27: Listening Test--Linear Growth Model with Age and Academic Status Interactions

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect       Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00       0.458174         0.090978        5.036     0.000
        MORE, B01      -0.901406         0.123594       -7.293     0.000
         AGE, B02      -0.023918         0.010262       -2.331     0.020
         UGR, B03       0.170578         0.112092        1.522     0.128
        UGXM, B04       0.108270         0.181319        0.597     0.550
        GRAD, B05       0.040756         0.134244        0.304     0.761
        GRXM, B06       0.173413         0.229100        0.757     0.449
 For   LINEAR slope, Π1
    INTRCPT2, B10       0.523732         0.062797        8.340     0.000
        MORE, B11      -0.369412         0.066112       -5.588     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect     Standard       Variance     df    Chi-square    P-value
                   Deviation      Component
 ----------------------------------------------------------------------------
 INTRCPT1, R0       0.42342        0.17929     192    489.16411     0.000
  level-1, E        0.49894        0.24894
```

Note that the interaction terms (UGXM and GRXM) represent more than one degree of freedom. In this situation, the t-tests used to evaluate the model's interactions' fixed effects are misleading. The point in adding the interaction terms is to establish the overall adequacy of the model not necessarily to compare interaction terms with arbitrarily chosen variables—in this case students studying English. The most appropriate method for determining model adequacy is change in explained variance (Equation 5). Unfortunately there is no significance test in HLM for change in explained variance. For this and subsequent models when interactions represent more than one degree of freedom, a change in explained variance greater than 2% will be considered significant. The reduction in variance resulting from adding the academic by more interactions is 8.0%. This represents a 2.4% reduction in variance compared to the model in Table 25. Prior to adopting this model, however, a model investigating the main effects of academic status is explored. If the main effects model represents a greater reduction in variance, it will be adopted in lieu of the interaction model. In this model, equation 12b is reduced to the following:

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{Age})_i + B_{03}\,(\text{Ugr})_i + B_{04}\,(\text{Grad})_i + r_{0i}.$$

Table 28 presents the results of this model.

Table 28: Listening Test-- Linear Growth Model with Age and Academic Status

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect        Coefficient   Standard Error  T-ratio    P-value
 ----------------------------------------------------------------------------
 For        INTRCPT1, Π0
    INTRCPT2, B00        0.431442        0.084603         5.100      0.000
        MORE, B01       -0.839138        0.096771        -8.671      0.000
         AGE, B02       -0.023268        0.010096        -2.305      0.021
         UGR, B03        0.214425        0.088350         2.427      0.015
        GRAD, B04        0.095392        0.116222         0.821      0.412
 For    LINEAR slope, Π1
    INTRCPT2, B10        0.523616        0.062793         8.339      0.000
        MORE, B11       -0.369460        0.066098        -5.590      0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect     Standard        Variance      df     Chi-square    P-value
                   Deviation       Component
 ----------------------------------------------------------------------------
 INTRCPT1, R0      0.42120         0.17741      194      491.64598     0.000
  level-1, E       0.49891         0.24891
```

In this model, the undergraduate student fixed effect has a significant t-value. Undergraduate students are predicted to have initial listening scores that are .214 logits greater than their other academic peers. The reduction in variance represented by this model is 9.0%. This is a 3.4% change in variance compared to the model in Table 25. Since the main effects model explains more variance than the interaction model, it is retained in lieu of the interaction model (Table 27).

The last set of characteristics to evaluate is students' language groups. Students from several unique language groups are included in the following analysis: Middle Eastern, Chinese, Japanese, and Korean students. There is also an amalgam of different languages put together in the category termed "other." This group represents students from a variety of countries and languages (European, South East Asian, etc). Because each individual (language) group within the “other” category is small, it is not possible to create additional separate groups. Equation 13 presents the interaction model for this analysis.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i} \text{ (linear term)}_i + e_{ti} \qquad (13a)$$

$$\Pi_{0i} = B_{00} + B_{01} \text{ (More)}_i + B_{02} \text{ (Age)}_i + B_{03} \text{ (ME)}_i + B_{04} \text{ (ME x More)}_i +$$
$$B_{05} \text{ (CHN)}_i + B_{06} \text{ (CHN x More)}_i + B_{07} \text{ (JPN)}_i +$$
$$B_{08} \text{ (JPN x More)}_i + B_{09} \text{ (KRN)}_i + B0_{10} \text{ (KRN x More)}_i +$$
$$r_{0i} \qquad (13b)$$

$$\Pi_{1i} = B_{10} + B_{11} \text{ (More)}_i \qquad (13c)$$

In Equation 13, ME represents Middle Eastern students; CHN represents Chinese students; JPN represents Japanese students, and KRN represents Korean students. The "x More" represents the language group by term of attendance interaction. Table 29 presents the results of this analysis.

Table 29: Listening Test--Linear Growth Model with Age and Language Group Interactions

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------------

 For        INTRCPT1, Π0
    INTRCPT2, B00        0.775636        0.140495         5.521     0.000
        MORE, B01       -1.156622        0.262054        -4.414     0.000
         AGE, B02       -0.019359        0.009882        -1.959     0.050
         UGR, B03        0.207167        0.087178         2.376     0.018
          ME, B04       -0.279302        0.205443        -1.360     0.174
        MEXM, B05        0.275300        0.359778         0.765     0.444
         CHN, B06       -0.464409        0.184743        -2.514     0.012
       CHNXM, B07        0.454921        0.358138         1.270     0.204
         JPN, B08       -0.354304        0.168658        -2.101     0.035
       JPNXM, B09        0.272908        0.313853         0.870     0.385
         KRN, B010       -0.378545        0.147140         -2.573     0.010
       KRNXM, B011        0.397678        0.282247          1.409     0.159
 For   LINEAR slope, Π1
    INTRCPT2, B10        0.523895        0.062738         8.351     0.000
        MORE, B11       -0.369561        0.066039        -5.596     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect      Standard       Variance      df    Chi-square    P-value
                    Deviation      Component
 ----------------------------------------------------------------------------

 INTRCPT1, R0       0.42033        0.17668       187     473.30026     0.000
  level-1, E        0.49846        0.24847
```

The intercept variance, R0 = 0.177, is slightly smaller than that of the model in Table 28. The addition of the interaction variables reduced variance by only 0.4%. Interaction predictors are removed from the model, and a subsequent analysis is conducted evaluating language groups' main effects. Results of this analysis are presented in Table 30.

Table 30: Listening Test--Linear Growth Model with Age and Language Group

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect     Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00      0.695274       0.126885        5.480     0.000
        MORE, B01     -0.833148       0.096334       -8.649     0.000
         AGE, B02     -0.021760       0.009352       -2.327     0.020
         UGR, B03      0.202848       0.085905        2.361     0.018
          ME, B04     -0.209174       0.168180       -1.244     0.214
         CHN, B05     -0.343561       0.156525       -2.195     0.028
         JPN, B06     -0.296511       0.138471       -2.141     0.032
         KRN, B07     -0.267927       0.124167       -2.158     0.031
 For   LINEAR slope, Π1
    INTRCPT2, B10      0.523841       0.062748        8.348     0.000
        MORE, B11     -0.369282       0.066047       -5.591     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect     Standard       Variance      df    Chi-square   P-value
                   Deviation      Component
 ----------------------------------------------------------------------
 INTRCPT1, R0      0.41724        0.17409      191     479.52032    0.000
  level-1, E       0.49855        0.24855
```

Students in the Chinese, Japanese, and Korean language groups have initial logits significantly lower than the "other" language group. Middle Eastern students do not seem to have significantly lower performance compared to the "other" group. Substantial variance is still unaccounted for in the intercept. Adding language groups into the HLM listening model reduces unexplained variance by 1.7%. Table 31 summarizes significant findings from the listening test analysis.

Table 31: Summary of Listening Test HLM Models Significant Findings

Significant Effects	Intercept (R_0)		Slope (R_1)	
	variance	Cumulative variance	Variance	Cumulative variance
Linear Model (Table 20)	0.363	---	0.018	---
Linear + More (Table 22)	0.242	33.3%	0.007	60.3%
Fixed Slope + More (Table 23)	0.195	---	Fixed	---
Fixed Slope + More & Age (Table 25)	0.184	5.6%	Fixed	---
Fixed Slope + More & Age & Academic Status (Table 28)	0.177	9.0%	Fixed	---
Fixed Slope + More & Age & Academic Status & Language Group (Table 30)	0.174	10.7%	Fixed	---

The linear growth model (analysis shown in Table 20) is the accepted model for listening test analyses. Adding the variable "More" to the linear model (Table 22) accounts for 33.3% of the variance in the intercept and 60.3% of the variance in the slope. The next model retains the "More" variable in both intercept and slope but fixes the slope's random effects to zero (Fixed Slope + More, Table 23). The fixed slope model now becomes the model for comparison. When the term "age" is added (Table 25), a 5.6% reduction in intercept unexplained variance is seen, i.e., age accounts for 5.6% of the variance in initial starting scores. Adding academic status—specifically undergraduate students—reduces unexplained variance by an additional 3.4%. Finally, adding the language group variables (Table 29) accounts for an additional 1.7% (total of 10.7%) reduction in unexplained variability of initial test scores.

To summarize, the linear growth model (compared to the quadratic growth model) most appropriately fits the listening test data. Students who stay more than one term at the ELC have significantly lower initial listening test scores. These students also have significantly lower linear growth trajectories. Older students have significantly lower initial listening test scores as well. Undergraduate students have higher initial listening test scores, and certain language groups (Chinese, Japanese and Korean) have significantly lower initial listening test scores. The next section describes the reading test analysis.

5.3 Modeling Reading Growth Trajectory

As with the listening test analysis, the first model examined in the reading test analysis is the linear model. Equation 14 presents this model.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\ (\text{linear term})_i + e_{ti} \qquad (14a)$$

$$\Pi_{0i} = B_{00} + r_{0i} \qquad (14b)$$

$$\Pi_{1i} = B_{10} + r_{1i} \qquad (14c)$$

This model is identical to that reported for the listening test except now the outcome variable Y_{ti} represents the reading test logit at time "t" for examinee "i." Table 32 reports the results of this analysis.

Table 32: Reading Test--Linear Growth Model

```
Final estimation of fixed effects:
------------------------------------------------------------------------
    Fixed Effect     Coefficient   Standard Error  T-ratio   P-value
------------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00      0.937900        0.056025       16.741     0.000
 For   LINEAR slope, Π1
    INTRCPT2, B10      0.556789        0.035275       15.784     0.000

 Final estimation of variance components:
------------------------------------------------------------------------
 Random Effect      Standard     Variance     df    Chi-square   P-value
                    Deviation     Component
------------------------------------------------------------------------
 INTRCPT1, R0        0.80866        0.65393   225   2597.77738    0.000
   LINEAR slope,R1 0.36620          0.13411   225    532.18409    0.000
  level-1, E         0.32130        0.10323


Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.

Tau (as correlations)
 INTRCPT1  1.000 -0.023
   LINEAR -0.023  1.000
```

The mean intercept (B00) and mean slope (B10) are 0.938 and 0.557 respectively. Both fixed effects have high t-ratios. Similar to the listening test model, the term LINEAR is centered in SPSS but with an average value of 0.75. This value is subtracted from each term variable; thus, the initial term value is -0.75 (0 - 0.75); the first term value is 0.25 (1 - 0.75), and so on. The predicted initial score for the reading test linear model is 0.938 + (-0.75 x 0.557) or 0.520. The predicted first term reading score is 1.077 (0.938 + (0.25 x 0.557). Predicted scores for other terms are calculated in a similar fashion. Random coefficients for the intercept and slope parameters are both significant at the p<.01 level, which indicates that substantial variation is not accounted for by this model. This suggests pursuit of more complex models. The correlation between the intercept and slope parameter is -0.023. There is practically no relationship between students' initial test scores and growth rates. This relationship is so small that further discussion of the correlation between the intercept and slope (for the linear model, at least) will be suspended.

Prior to continuing with the linear model, a quadratic model is explored. The equation for this model is presented below.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\text{ (linear term)}_i + \Pi_{2i}\text{ (quadratic term)}_i + e_{ti} \qquad (15a)$$

$$\Pi_{0i} = B_{00} + r_{0i} \qquad (15b)$$

$$\Pi_{1i} = B_{10} + r_{1i} \qquad (15c)$$

$$\Pi_{2i} = B_{20} + r_{2i} \qquad (15d)$$

The new term in this equation is "Π_{2i} (quadratic term)$_i$" which is the linear term squared. Table 33 presents the results of this model's analysis.

Table 33: Reading Test--Quadratic Growth Model

```
Final estimation of fixed effects:
----------------------------------------------------------------------
   Fixed Effect     Coefficient   Standard Error  T-ratio   P-value
----------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00     0.960634       0.058430      16.441     0.000
 For   LINEAR slope, Π1
    INTRCPT2, B10     0.533966       0.034703      15.387     0.000
 For     QUAD slope, Π2
    INTRCPT2, B20    -0.090898       0.032329      -2.812     0.005

 Final estimation of variance components:
 ---------------------------------------------------------------------
 Random Effect   Standard     Variance     df    Chi-square   P-value
                 Deviation    Component
 ---------------------------------------------------------------------
 INTRCPT1, R0      0.81969       0.67189    34     438.36683    0.000
   LINEAR slope,R1 0.33692       0.11351    34      81.12967    0.000
     QUAD slope,R2 0.09093       0.00827    34      43.10633    0.136
  level-1, E       0.33603       0.11292

Note: The chi-square statistics reported above are based on only 35 of
228 units that had sufficient data for computation.


 Statistics for current covariance components model
 --------------------------------------------------
 Deviance =    1066.19783
 Number of estimated parameters =    7

Tau (as correlations)
 INTRCPT1  1.000 -0.140 -0.736
   LINEAR -0.140  1.000  0.271
     QUAD -0.736  0.271  1.000
```

In the quadratic model, all fixed effects are significant. The quadratic component is negative (B20 = -0.091), which indicates that across terms, students' growth rates become less steep. Both the intercept and slope random effects are significant; however, the quadratic random component is not (chi-square = 43.106, p= 0.136). This indicates that the quadratic component of the model cannot support and does not require further fixed effects. Three sets of correlation coefficients are displayed in the quadratic model:

an intercept to linear slope correlation (-0.140), an intercept to quadratic component correlation (-0.736) and a linear slope to quadratic component correlation (0.271). The intercept to slope correlation is higher and more negative than the linear model. The intercept to quadratic component correlation, which is relatively high, implies that the lower a student's initial test score the less their score tapers off over terms. Since this is an unconditional model, i.e., no predictors in the fixed effects, no variance components are calculated.

A decision is now required. Should the unconditional linear or quadratic model be adopted? As stated earlier, there is theoretical appeal to adopting the quadratic model. However, several indicators suggest that the linear model is a better fit for the data. First, Figure 2b, in Chapter 4, portrays a comparison between reading test score trajectories of students staying one term at the ELC and those of students who stayed longer. There is clearly a linear-like trajectory for both groups of students. Second, level-1 variance is smaller for the linear model ($v(e_i) = 1.0912$) than for the quadratic model ($v(e_i) = 1.1902$). Third, the number of iterations required to converge the models is substantially smaller for the linear model (14 iteration cycles) than for the quadratic model (867 iteration cycles). Fourth, there are very few time points from which to generate a strong non-linear model--only a maximum of 5 time points with a majority of students only having two. Finally, only 35 students had more than 2 data points from which to calculate the quadratic model random effects. That means that of the 228 students in the reading sample, only 15% would be used to evaluate the random variance components of future models. For these reasons, the quadratic model is rejected, and the linear model is adopted for further HLM analysis.

The next linear model adds the term "More." Equation 16 displays this model.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti}\,, \tag{16a}$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + r_{0i}\,, \tag{16b}$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + r_{1i}\,. \tag{16c}$$

Table 34 presents the results of this analysis.

Table 34: Reading Test--Linear Model with Person Variable "More"

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect     Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00      1.195000       0.058484       20.433     0.000
        MORE, B01     -0.884070       0.109766       -8.054     0.000
 For   LINEAR slope, Π1
    INTRCPT2, B10      0.627795       0.046291       13.562     0.000
        MORE, B11     -0.122839       0.071061       -1.729     0.083

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect     Standard    Variance     df    Chi-square   P-value
                   Deviation   Component
 ----------------------------------------------------------------------
 INTRCPT1, R0        0.69833     0.48767   224    1901.70433    0.000
   LINEAR slope,R1 0.35322     0.12476   224     494.82768    0.000
  level-1, E        0.32779     0.10745

Statistics for current covariance components model
 --------------------------------------------------
 Deviance =    1010.12305
 Number of estimated parameters =    4
```

The term "More" has a high t-ratio in the intercept portion of the model. However, it is not significant at the p<.05 level in the slope segment (p = .083). This could support removing this component from the slope portion of the model. However, future models incorporate interaction effects between other person variables and "More." Bryk and Raudenbush (1992) express a note of caution when removing a variable from one portion of a model but retaining it another. This is especially true with interactions. They argue that misspecifications error and bias can make interpretations difficult--this occurs most frequently when variables are interrelated (p.215). Because of this, the "More" variable will be retained in the slope segment of the model.

Both random effect parameter estimates have high chi-square values and are significant at the $p<.01$ level. This means that substantial variation has yet to be explained in the intercept and slope segments of the model. Future exploration of more complex models seems warranted.

The "More" variable, when added to the intercept portion of the model, accounts for 25.4% of the variability (using equation 5 and variance estimates from Tables 32 and 34). The variance accounted by More in the slope is substantially less at only 7.0%.

The next model investigates whether age by More interactions exist in reading growth trajectories. Equation 17 illustrates this model.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti} \tag{17a}$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{Age})_i + B_{01}\,(\text{Age x More})_i + r_{0i} \tag{17b}$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + B_{12}\,(\text{Age})_i + B_{13}\,(\text{Age x More})_i + r_{1i} \tag{17c}$$

Table 35: Reading Test—Linear Model with More, Age and Age by More Interaction

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect     Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For       INTRCPT1, Π0
    INTRCPT2, B00      1.190049       0.058557       20.323     0.000
        MORE, B01     -0.906025       0.111017       -8.161     0.000
         AGE, B02      0.013612       0.011805        1.153     0.249
       AGEXM, B03     -0.033750       0.024558       -1.374     0.169
 For   LINEAR slope, Π1
    INTRCPT2, B10      0.630872       0.045273       13.935     0.000
        MORE, B11     -0.079436       0.068342       -1.162     0.246
         AGE, B12     -0.010268       0.009154       -1.122     0.262
       AGEXM, B13      0.060410       0.015055        4.013     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect      Standard       Variance      df    Chi-square    P-value
                    Deviation      Component
 ----------------------------------------------------------------------
 INTRCPT1, R0         0.69505        0.48310    222    1762.89070     0.000
   LINEAR slope,R1  0.30791        0.09481    222     405.00759     0.000
  level-1, E          0.33813        0.11433
```

The intercept (B00) and More (B01) parameter estimates in the intercept portion of the model again have significant t-ratios. Initial reading scores seem not to be significantly different between students of different ages. For the slope parameter, the intercept (B10) and the age by More interaction (B13) are significant fixed effects. This finding indicates that older students who stay more than one term at the ELC have significantly higher growth rates. In the slope portion of the model, both the More and age main effects do not have significantly high t-ratios. The intercept and slope's random effects have significantly high chi-square values; thus, substantial variation is still unexplained by this model. Further investigation of fixed effects seems warranted. For parsimony, a further reduced model is represented in Equation 18.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti}\,, \qquad (18a)$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + r_{0i}\,, \qquad (18b)$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + B_{12}\,(\text{Age})_i + B_{13}\,(\text{Age x More})_i + r_{1i}\,. \qquad (18c)$$

In this new model, the age by More interaction is removed from the intercept portion of the model, and the age, More and age by More interaction fixed effects are retained in the slope portion of the model. Table 36 presents the results of this new model.

Table 36: Reading Test—Linear Model with More and Age by More Interaction

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------------
 For        INTRCPT1, Π0
    INTRCPT2, B00       1.193770         0.058499        20.407     0.000
        MORE, B01      -0.893333         0.109645        -8.148     0.000
 For    LINEAR slope, Π1
    INTRCPT2, B10       0.630813         0.045258        13.938     0.000
        MORE, B11      -0.083521         0.068285        -1.223     0.222
         AGE, B12      -0.010882         0.009135        -1.191     0.234
       AGEXM, B13       0.057568         0.014614         3.939     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect     Standard       Variance     df     Chi-square    P-value
                   Deviation      Component
 ----------------------------------------------------------------------------
 INTRCPT1, R0        0.69576        0.48408   224     1797.41036     0.000
   LINEAR slope,R1  0.30986        0.09601   222      408.33857     0.000
  level-1, E         0.33710        0.11364

Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.
```

In this model, the intercept's fixed effect "More" (B01) has a high t-ratio; likewise, the slope's age by More interaction (B13) has a high t-ratio. The variance in the intercept explained by this model is 26.0%. This is only 0.6% higher than that of the model incorporating the More variable (Equation 16). However, the amount of variance explained in the slope parameter is substantially higher than previous models. Recall that only 7.0% of slope variance is accounted for by the model portrayed in Equation 16. Using Equation 5 and values from Tables 32 and 36, we see that the model displayed in Equation 18 accounts for 28.4% of slope parameter variance. Thus retention of the More variable in the intercept and the age by More interaction in the slope for future models is adopted.

The next model, shown below, investigates whether a sex-by-More interaction exists.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti} \qquad (19a)$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{Sex})_i + B_{01}\,(\text{Sex x More})_i + r_{0i} \qquad (19b)$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + B_{12}\,(\text{Sex})_i + B_{13}\,(\text{Sex x More})_i +$$

$$B_{14}\,(\text{Age})_i + B_{15}\,(\text{Age x More})_i + r_{1i} \qquad (19c)$$

Table 37 presents the results of this analysis.

Table 37: Reading Test—Linear Model with More, Age by More Interaction, and Sex by More Interaction

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------

 For       INTRCPT1, Π0
    INTRCPT2, B00       1.175411        0.077661        15.135     0.000
        MORE, B01      -0.991126        0.148655        -6.667     0.000
         SEX, B02       0.043421        0.117875         0.368     0.712
         SXM, B03       0.206470        0.219852         0.939     0.348
 For   LINEAR slope, Π1
    INTRCPT2, B10       0.711617        0.060216        11.818     0.000
        MORE, B11      -0.196331        0.091610        -2.143     0.032
         SEX, B12      -0.184771        0.091197        -2.026     0.042
         SXM, B13       0.254044        0.135235         1.879     0.060
         AGE, B14      -0.012800        0.009142        -1.400     0.161
       AGEXM, B15       0.060970        0.014633         4.167     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect     Standard      Variance     df    Chi-square    P-value
                   Deviation     Component
 ----------------------------------------------------------------------
 INTRCPT1, R0       0.69517      0.48326   222    1778.51538    0.000
   LINEAR slope,R1 0.30789      0.09480   220     404.20237    0.000
  level-1, E        0.33569      0.11269

Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.
```

The sex by More interaction effect (B03) is not significant in the intercept portion of the model. The sex by More interaction effect (B13) is not significant at the p<.05 level in the slope portion of the model. However, one might consider this variable to be marginally significant. But the amount of explained variance represented by adding the sex by More interaction is inconsequential (only 0.7%). Thus, the sex by More interaction will be discarded from future models. While the sex main effect t-value is high in the slope portion of this model, interpretation is problematic since an interaction term is present. A subsequent model is tested removing the sex by More interaction terms from the intercept and slope portions of the model. The sex main effect is not significant in either the intercept or slope portion of this new model. Taken together, there are no significant differences in initial reading test scores or growth trajectories between sexes or between sexes who stay at the ELC one term or longer. Both random effect parameter estimates have significantly high chi-square values, indicating that substantial variation still exists in the intercept and slope portions of the model. Again, this justifies further exploration.

The next model adds variables related to academic status. This model is displayed in the equation below.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti} \tag{20a}$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{Ugrad})_i + B_{03}\,(\text{Ugrad x More})_i + B_{04}\,(\text{Grad})_i + B_{05}\,(\text{Grad x More})_i + r_{0i} \tag{20b}$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + B_{12}\,(\text{Age})_i + B_{13}\,(\text{Age x More})_i + B_{14}\,(\text{Ugrad})_i + B_{15}\,(\text{Ugrad x More})_i + B_{16}\,(\text{Grad})_i + B_{17}\,(\text{Grad x More})_i + r_{1i} \tag{20c}$$

Table 38: Reading Test-- Linear Model with More, Age by More Interaction and Academic Status by More Interaction

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio    P-value
 ----------------------------------------------------------------------------

 For        INTRCPT1, Π0
    INTRCPT2, B00       1.104267         0.096713         11.418      0.000
        MORE, B01      -0.886698         0.164787         -5.381      0.000
       UGRAD, B02       0.121449         0.134446          0.903      0.367
        UGXM, B03       0.150214         0.241494          0.622      0.534
        GRAD, B04       0.170648         0.152743          1.117      0.264
        GDXM, B05      -0.330460         0.321599         -1.028      0.305
 For    LINEAR slope, Π1
    INTRCPT2, B10       0.657609         0.075828          8.672      0.000
        MORE, B11      -0.159205         0.105221         -1.513      0.130
         AGE, B12      -0.006967         0.010303         -0.676      0.499
       AGEXM, B13       0.050464         0.017028          2.964      0.004
       UGRAD, B14      -0.004196         0.106225         -0.039      0.969
        UGXM, B15       0.051779         0.152852          0.339      0.735
        GRAD, B16      -0.106895         0.125790         -0.850      0.396
        GDXM, B17       0.297912         0.212720          1.400      0.161

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect      Standard        Variance      df     Chi-square    P-value
                    Deviation       Component
 ----------------------------------------------------------------------------

 INTRCPT1, R0         0.69333       0.48071      220     1711.62269     0.000
   LINEAR slope,R1  0.30701         0.09426      218      391.16097     0.000
  level-1, E          0.34064       0.11604

Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.
```

Recall that when interaction effects represent more than one degree of freedom model adequacy is determined by explanatory variance. A 2% or greater amount of explanatory variance is considered significant when interpreting interaction effects in this situation. The amount of explained variance in this model's intercept is 26.5%, and in this model's slope, it is 29.7%. Table 36's model (the last model with significant predictors) has intercept and slope variances of 26.0% and 28.4% respectively. Both variances change less than 2%. Hence, academic status by More interactions are not considered significant.

Another model is run removing academic by More interaction effects from the intercept and slope portions of the model. In this new model, none of the academic main effects are significant.

Substantial variation still exists in the random effects portion of the model. Both random effect parameters have significantly high chi-square values. The non-significant findings in fixed effects portions of the model preclude retention of academic variables.

The next and final model incorporates language group and is portrayed in the following equation.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i} \text{ (linear term)}_i + e_{ti}\text{ ,} \tag{21a}$$

$$\begin{aligned}\Pi_{0i} = {} & B_{00} + B_{01} \text{ (More)}_i + B_{02} \text{ (ME)}_i + B_{03} \text{ (ME x More)}_i + B_{04} \text{ (CHN)}_i + \\ & B_{05} \text{ (CHN x More)}_i + B_{06} \text{ (JPN)}_i + B_{07} \text{ (JPN x More)}_i + B_{08} \text{ (KRN)}_i + \\ & B_{09} \text{ (KRN x More)}_i + r_{0i}\text{ ,}\end{aligned} \tag{21b}$$

$$\begin{aligned}\Pi_{1i} = {} & B_{10} + B_{11} \text{ (More)}_i + B_{12} \text{ (Age)}_i + B_{13} \text{ (Age x More)}_i + \\ & B_{14} \text{ (ME)}_i + B_{15} \text{ (ME x More)}_i + B_{16} \text{ (CHN)}_i + B_{17} \text{ (CHN x More)}_i + \\ & B_{18} \text{ (JPN)}_i + B_{19} \text{ (JPN x More)}_i + B_{110} \text{ (KRN)}_i + \\ & B_{111} \text{ (KRN x More)}_i + r_{1i}\text{ .}\end{aligned} \tag{21c}$$

Several language groups are represented in this model: Middle Eastern (ME), Chinese (CHN), Japanese (JPN), and Korean (KRN). There is also another group, which is essentially an amalgam of language groups (e.g., European and South East Asian)—termed "Other." Each language group is denoted by a dummy variable, and the intercept for both initial status and linear growth components represents values for the "Other" language group. Table 39 displays the results of an HLM analysis of this model.

Table 39: Reading Test-- Linear Model with More, Age by More Interaction and Language Group by More Interactions

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------------
 For        INTRCPT1, Π0
    INTRCPT2, B00       1.396021        0.125638        11.111     0.000
        MORE, B01      -1.089622        0.451804        -2.412     0.016
          ME, B02      -1.036685        0.250202        -4.143     0.000
        MEXM, B03       1.188360        0.579372         2.051     0.040
         CHN, B04      -0.122613        0.187698        -0.653     0.513
       CHNXM, B05      -0.386730        0.570274        -0.678     0.497
         JPN, B06      -0.295885        0.183376        -1.614     0.106
       JPNXM, B07      -0.026254        0.500705        -0.052     0.959
         KRN, B08      -0.154584        0.154693        -0.999     0.318
       KRNXM, B09       0.365143        0.477357         0.765     0.444
 For    LINEAR slope, Π1
    INTRCPT2, B10       0.754984        0.101475         7.440     0.000
        MORE, B11      -0.358831        0.254380        -1.411     0.158
         AGE, B12      -0.012888        0.009251        -1.393     0.164
       AGEXM, B13       0.065180        0.015992         4.076     0.000
          ME, B14      -0.045771        0.198113        -0.231     0.817
        MEXM, B15      -0.049509        0.343837        -0.144     0.886
         CHN, B16      -0.198380        0.151312        -1.311     0.190
       CHNXM, B17       0.644898        0.333140         1.936     0.052
         JPN, B18      -0.133375        0.146374        -0.911     0.363
       JPNXM, B19       0.319339        0.301150         1.060     0.289
         KRN, B110      -0.164276        0.125168        -1.312     0.190
       KRNXM, B111       0.320119        0.277029         1.156     0.248

 Final estimation of variance components:
 ----------------------------------------------------------------------------
 Random Effect      Standard       Variance      df     Chi-square    P-value
                    Deviation      Component
 ----------------------------------------------------------------------------
 INTRCPT1, R0         0.66645        0.44416     216    1647.35263     0.000
   LINEAR slope,R1 0.31072           0.09655     214     391.53948     0.000
  level-1, E          0.33575        0.11273

Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.
```

This model is assessing interaction terms that have more than one degree of freedom. The amount of variance accounted for in the intercept portion of this model is 32.1%, and the amount accounted for in the slope portion of the model is 28.0%. Using variance estimates from Table 36 as the comparison, this model's intercept increases explanatory variance by 5.9%. Conversely, this models' slope has a 0.4% reduction in explanatory variance. This suggests that the interaction terms should be retained in the intercept portion of the model and removed from the slope portion of the model. In this reduced model, equation 21c becomes:

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + B_{12}\,(\text{Age})_i + B_{13}\,(\text{Age x More})_i + r_{1i}\,.$$

Results of variance estimates for this reduced model are R0=0.479, and R1=0.096. This model's variance estimates represent only a 0.8% gain in the intercept and a –0.3% loss in the slope. We now have a quandary. When interaction terms are added to both the intercept and slope, substantial variance is found in the intercept. When the interactions are removed from the slope portion of the model, variance estimates for the intercept are not significant. These conflicting findings suggest that a model with main effects may be appropriate. This model is represented by the following equation.

$$Y_{ti} = \Pi_{0i} + \Pi_{1i}\,(\text{linear term})_i + e_{ti}\,, \qquad (22a)$$

$$\Pi_{0i} = B_{00} + B_{01}\,(\text{More})_i + B_{02}\,(\text{ME})_i + B_{03}\,(\text{CHN})_i + B_{04}\,(\text{JPN})_i + B_{05}\,(\text{KRN})_i + r_{0i}\,, \qquad (22b)$$

$$\Pi_{1i} = B_{10} + B_{11}\,(\text{More})_i + B_{12}\,(\text{Age})_i + B_{13}\,(\text{Age x More})_i + B_{14}\,(\text{ME})_i + (\text{CHN})_i + B_{16}\,(\text{JPN})_i + B_{17}\,(\text{KRN})_i + r_{1i}\,. \qquad (22c)$$

Table 40: Reading Test-- Linear Model with More, Age by More Interaction and Language Group

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio    P-value
 ----------------------------------------------------------------------------
 For        INTRCPT1, B0
    INTRCPT2, G00       1.362225        0.122237        11.144      0.000
        MORE, G01      -0.846037        0.110592        -7.650      0.000
          ME, G02      -0.689003        0.215322        -3.200      0.002
         CHN, G03      -0.170651        0.177565        -0.961      0.337
         JPN, G04      -0.365307        0.164875        -2.216      0.027
         KRN, G05      -0.075507        0.145320        -0.520      0.603
 For    LINEAR slope, B1
    INTRCPT2, G10       0.698982        0.093858         7.447      0.000
        MORE, G11      -0.070829        0.071084        -0.996      0.319
         AGE, G12      -0.011317        0.009258        -1.222      0.222
       AGEXM, G13       0.056545        0.015011         3.767      0.000
          ME, G14      -0.111258        0.151587        -0.734      0.463
         CHN, G15      -0.065815        0.133349        -0.494      0.621
         JPN, G16      -0.085639        0.122087        -0.701      0.483
         KRN, G17      -0.088970        0.108802        -0.818      0.414

 Final estimation of variance components:
 -----------------------------------------------------------------------------
 Random Effect       Standard       Variance      df    Chi-square    P-value
                     Deviation      Component
 -----------------------------------------------------------------------------
 INTRCPT1,  U0        0.67781       0.45943      220    1707.25576     0.000
   LINEAR slope,U1    0.31606       0.09990      218     409.17035     0.000
  level-1,  R         0.33615       0.11300

Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.

 Statistics for current covariance components model
 --------------------------------------------------
 Deviance =    1028.87379
 Number of estimated parameters =     4
```

The Middle Eastern (ME) and Japanese (JPN) language groups have significant t-values in the intercept portion of this model. No language group exhibits significant t-values in the slope portion of the model. Variance estimates for this model are 0.459 for the intercept and 0.100 for the slope. This represents a 3.7% increase in the intercept portion of the model and a 0.9% loss in the slope portion of the model. These findings suggest removing language group from the slope portion of the model and retaining the Middle Eastern and Japanese language groups in the intercept portion of the model. Table 41 presents the results of this new analysis.

Table 41: Reading Test--Linear Model with More, ME and JPN Group in the Intercept and Age by More Interaction in the Slope

```
Final estimation of fixed effects:
 ----------------------------------------------------------------------
    Fixed Effect      Coefficient   Standard Error  T-ratio   P-value
 ----------------------------------------------------------------------
 For       INTRCPT1, B0
    INTRCPT2, G00      1.285650        0.062720      20.498     0.000
        MORE, G01     -0.849624        0.107506      -7.903     0.000
          ME, G02     -0.615429        0.182865      -3.365     0.001
         JPN, G03     -0.287001        0.119977      -2.392     0.017
 For   LINEAR slope, B1
    INTRCPT2, G10      0.630327        0.045266      13.925     0.000
        MORE, G11     -0.085374        0.068102      -1.254     0.210
         AGE, G12     -0.010929        0.009138      -1.196     0.232
       AGEXM, G13      0.057423        0.014540       3.949     0.000

 Final estimation of variance components:
 ----------------------------------------------------------------------
 Random Effect      Standard       Variance      df    Chi-square    P-value
                    Deviation      Component
 ---------------------------------------------------------------------
 INTRCPT1, U0         0.67414      0.45447     222    1690.14458    0.000
   LINEAR slope,U1  0.30685        0.09416     222     404.70047    0.000
  level-1, R          0.33854      0.11461

Note: The chi-square statistics reported above are based on only 226 of
228 units that had sufficient data for computation.
```

In this final reading test model, the Middle Eastern and Japanese language groups have significant t-values in the intercept. This model (R0=0.454 and R0=0.094) accounts for 30.5% of the variance in the intercept and 29.8% of the variance in the slope. This represents a 4.5% increase explained variance in the intercept and a 1.4% increase in the slope compared to the estimates found in Table 36. There is still substantial variance not accounted for in the model—as represented by the random effects variance components high chi-square values. Table 42 presents major findings from the reading test analysis.

Table 42: Summary of Reading Test HLM Models Significant Findings

Significant Effects	Intercept (R_0)		Slope (R_1)	
	variance	Cumulative variance	variance	Cumulative variance
Linear Model (Table 32)	0.654	---	0.134	---
Linear + More (Table 34)	0.488	25.4%	0.125	7.0%
Linear + More + Age by More (Table 36)	0.484	26.0%	0.096	28.4%
Linear + More + Age + ME & JPN Lang. Group (Table 41)	0.454	30.5%	0.094	29.8%

To sum, the linear model fits the reading test data best. The addition of the More variable accounts for 25.4% of the intercept variation and accounts for 7.0% of the variation in the slope. The addition of an age by More interaction in the slope portion of the model dramatically increases explained variance—from 7.0% to 28.4%. The addition of Middle Eastern and Japanese language groups into the intercept portion of the model increases variance by 4.5% in the intercept and 1.4% in the slope.

The next chapter takes the findings reported here for listening and reading comprehension tests and addresses key research questions.

CHAPTER 6: RESEARCH QUESTIONS AND DISCUSSION

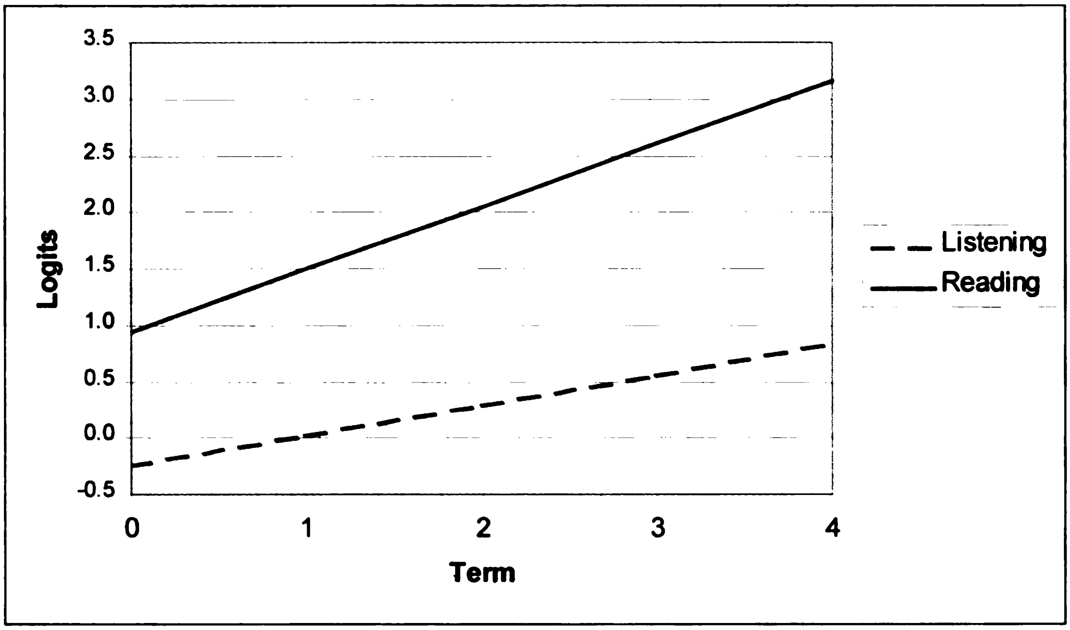
6.1 Research Questions

The goal of the research reported here was to investigate the growth characteristics of listening and reading portions of the English Language Center's placement test. Three research questions drove this study. The following section presents each question and addresses major findings.

Research question 1: What is the nature of growth trajectories on the ELCPT's listening and reading comprehension subtests?

As mentioned in Chapter 2, few studies have investigated the nature of language test growth trajectories. Several researchers have argued that second language acquisition is non-linear; thus, one might expect language tests' growth trajectories to reflect this non-linear trend. Both linear and non-linear models are investigated here. In both the listening and reading comprehension test analyses the linear and non-linear models fit the data. There is some evidence--albeit limited--that students at lower ability levels reflect quadratic-like growth tendencies. However, the limited amount of low-level students' data prevents a thorough exploration of quadratic growth models. The linear growth model was the most parsimonious for both test data sets. Figure 3 presents the fitted linear model for each examination. While both examinations are placed upon the same graph, one should not be assume that they are on the same scale. Each test is scaled separately. Thus, it is inappropriate to conclude that students score higher on the reading test than on the listening test. The main purpose for placing both trajectories upon the same figure is to illustrate the differences in predicted slopes (rates of growth).

Figure 3: Predicted English Language Center Listening and Reading Comprehension Test Growth Trajectories--Best Fit



From Figure 3, it is clear that the reading test's growth trajectory is steeper than the listening test's trajectory.

Using correlations between initial status and linear growth parameter estimates (Tables 20 and 32), this study finds that students with low listening test scores tend to have lower growth rates. Students with high initial listening test scores tend to have higher (more advantageous) growth rates. On the reading test, there is very little relationship between students' initial reading scores and linear growth rates.

Research Question 2: What demographic factors affect students' growth on the ELCPT's listening and reading comprehension tests?

Five demographic factors are explored here: length of stay at the ELC (More), sex, age, academic status, and language group. Each factor is inserted into an HLM model for each test. In addition interactions between these factors and the variable More are investigated. Table 43 presents the significant findings for each test.

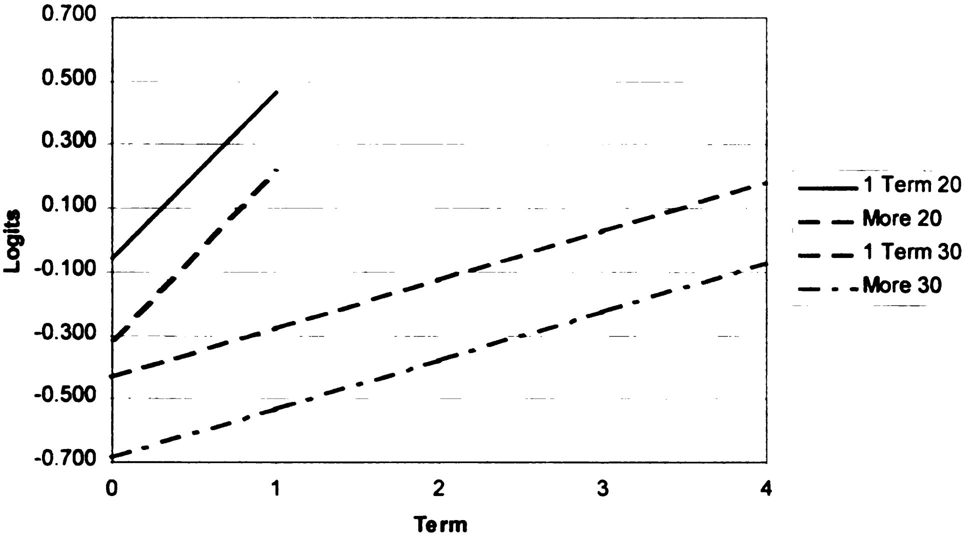
Table 43: Significant Demographic Findings and Their Variances on ELCPT Listening and Reading Comprehension Tests

	Listening		Reading	
Variables	Intercept (A_0)	Slope (A_1)	Intercept (A_0)	Slope (A_1)
More	33.3%	60.3%	25.4%	7.0%
Sex	n.s.	n.s.	n.s.	n.s.
Sex by More	n.s.	n.s.	n.s.	n.s.
Age	5.6%	n.s.	n.s.	n.s.
Age by More	n.s.		26.0%	28.4%
Academic status	9.0%	n.s.	n.s.	n.s.
Academic status by More	n.s.	n.s.	n.s.	n.s.
Language Group	10.7%	n.s.	30.5%	29.8%
Language Group by More	n.s.	n.s.	n.s.	n.s

n.s. = not significant

On both the listening and reading comprehension test HLM analyses, the variables More, and language group have significant findings. Students who stay more than one term at the ELC have significantly lower initial scores and lower growth rates on the listening comprehension test. Significantly lower initial listening scores for students staying longer than one term is not a surprising finding, since their retention at the ELC is primarily due to limited English proficiency. However, significantly lower growth trajectories for these students (on the listening test) are not an expected finding. Essentially, students entering the ELC with lower initial listening scores do not progress at the same rate as students with higher initial scores--further substantiated by the high correlation observed between initial status and growth rate mentioned above. Age also plays a significant role in initial listening scores. The older a student the more likely she or he is to have lower initial listening scores. Undergraduate students tend to have better initial listening test scores, and finally, Chinese, Japanese and Korean students have significantly lower initial listening scores. For the listening test, the variables that manifested the largest explanatory effects are More and age. Figure 4 graphically displays students' predicted growth trajectories using the estimates found in Table 25. Table 25's analysis depicts growth based upon the variables More and age. This model is chosen to display because these two variables represent a majority of listening test explanatory variance.

Figure 4: Predicted Listening Comprehension Growth Trajectories for 20 year-old and 30 year-old Students Staying at the ELC for One and More than One Term

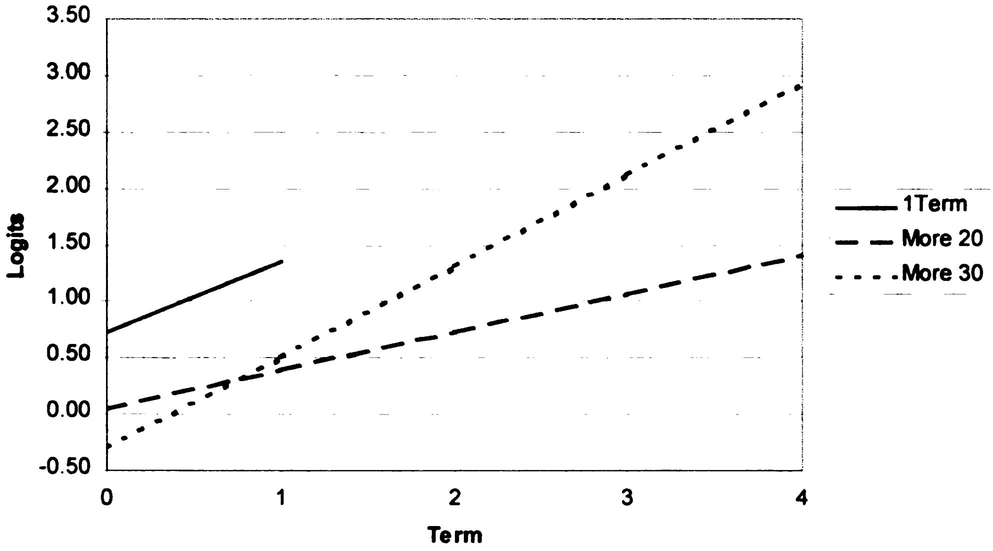


Notice the dramatic differences in growth trajectories for students staying one term (1 Term 20, 30) and those staying more than one term (More 20, 30). Notice also the large differences in initial scores between 20 year-olds and 30 year-olds.

For the reading comprehension test, the variable More has a significant effect upon students' initial reading scores; however, when used alone it has no import on students' growth rates. As with the listening test, age plays a significant role in growth on the reading test. There are no main effects for age in either initial score or growth rate. However, there is a significant interaction between age and More. That is, older students who stay at the ELC for more than one term have significantly higher growth rates. Similar to findings for the listening analysis, language groups have a significant effect--primarily in students' initial reading test scores. Middle Eastern and Japanese students have significantly lower initial reading test scores compared to other language groups.

However, a majority of the explained variance in demographic variables for the reading test is carried by the More and age by More variables. Figure 5 portrays this relationship.

Figure 5: Predicted Reading Comprehension Growth Trajectories for Students staying One Term and 20 and 30 year-old Students Staying More than One Term



The More by age interaction is clearly seen in the trajectories of those students staying more than one term at the ELC (More 20 and More 30). Older students fare better on the reading comprehension test. Like the listening test, students staying more than one term have substantially lower initial reading scores than their one term classmates.

To summarize, five demographic variables exhibit significant effects upon students' listening and reading comprehension growth trajectories: length of stay at the ELC (More), age, age by More interaction, academic group (undergraduates on the listening test) and language group. The other investigated variables do not seem to affect students' initial test scores or growth trajectories. Of the significant variables, two have the largest effect on initial scores and growth trajectories: More and age. Specific language groups (e.g., Chinese, Japanese and Korean students on the listening test and Middle Eastern and Japanese students on the reading test) have significantly different initial test scores, but the overall effects of their group membership on explanatory adequacy of HLM models is not great. While significant language group differences exist, they do not contribute greatly to HLM models, and thus strong inferences about differences in language group effects is not appropriate.

Research Question 3: What inferences might be made regarding the use of the ELCPT's listening and reading comprehension tests, based upon the discovered student growth trajectories and demographic factors?

Language tests are used in most, if not all, university level language programs. Brown (1996) suggests that most language tests are used with four particular purposes in mind: proficiency, placement, achievement and diagnostic. The examinations used in this study are used for placement and achievement purposes. Incoming international students not meeting MSU's English language admission requirement take the ELCPT to place into or out of the English Language Center classes. Once in the program, students take the test to advance to higher levels or to exit the English Language Center. The findings of this work provide guidance on several key decisions made with the listening and reading portions of the ELCPT.

The ELCPT listening and reading comprehension tests are both used for placement and achievement purposes. Does it seem reasonable to use these examinations for placement and achievement purposes? What might one expect to see from a placement test or an achievement test?

The first and most important issue in answering the above questions is whether a test is valid and reliable for the decision being made. Test analyses presented in Chapter 4 provide evidence that the tests used in this study are reliable. The factor analytic studies conducted for the IRT analysis offer evidence of the tests' construct validity. Assuming that each exam meets basic criteria for reliability and validity, what evidence might be amassed to justify the use of the ELCPT listening and reading tests as placement and/or achievement measures? One piece of evidence suggesting appropriate placement test uses is the significant differences between students who stay at the ELC one term and those who stay longer. On both listening and reading comprehension tests, significant differences exist in initial test scores of students who stay one term and those who stay longer. These differences suggest, at least partly, that tests are discriminating between at least two levels of language ability. The argument is that if these tests did not discriminate between ability levels there would be little difference between students' initial scores based on the length of stay at the ELC. Further exploration of different placement levels that might be uncovered on this exam is limited by this study's small sample size. Nonetheless, there seems to be reasonable evidence that the tests' use as a placement tool is merited.

Fulcher (1997) investigated the psychometric characteristics of a British university placement test. He discovered in his investigation that little research had been done investigating gain characteristics of language exams--specifically placement tests. He correctly surmised that knowledge of these gain characteristics would assist program administrators, teachers and students in predicting potential times required to finish language coursework. Since the ELCPT is used as an achievement measure, it is possible to explore and present potential timelines needed--based upon performance on this exam--to complete ELC course requirements. Table 44 displays predicted listening and reading comprehension test logits based upon HLM models, specifically between students who stay one term and students who stay longer than one term at the ELC.

Table 44: Comparison Between Predicted Students' Scores on Listening and Reading Comprehension Tests and Students Staying One or More Terms at the ELC

	Listening Comprehension Test		Reading Comprehension Test	
Term	One Term	Two or More*	One Term	Two or More*
Initial Score	-0.156	-0.531	0.721	-0.173
1st Term	0.367	-0.377	1.351	0.375
2nd Term		-0.224		0.922
3rd Term		-0.070		1.469
4th Term		0.083		2.016

*represents predicted growth rate of average aged student

On the listening comprehension test, students staying one term have a predicted average 1st term logit estimate of 0.367. The 1st term logit estimate for the reading test is 1.351. If we assume that to exit the ELC, a student must obtain a logit score equal to or greater than 0.367 on the listening test and 1.351 on the reading test, we would predict that it would take more-than-one-term students 3 semesters to obtain a reading score high enough to exit the program. Similarly, it would take more-than-one-term listening students more than four terms to obtain a score high enough to clear the ELC. In fact, it would take six semesters for the average more-than-one-term student to obtain a high enough listening test score. Using the listening and reading tests' estimates presented in Tables 25 and 36, a predicted average completion time in terms (assuming that reaching the predicted average logit for one-term students is sufficient) is calculated and presented by age.

Figure 6a: Predicted Average Completion Time for Students Staying More Than One Term by Age Group on the Listening Comprehension Test

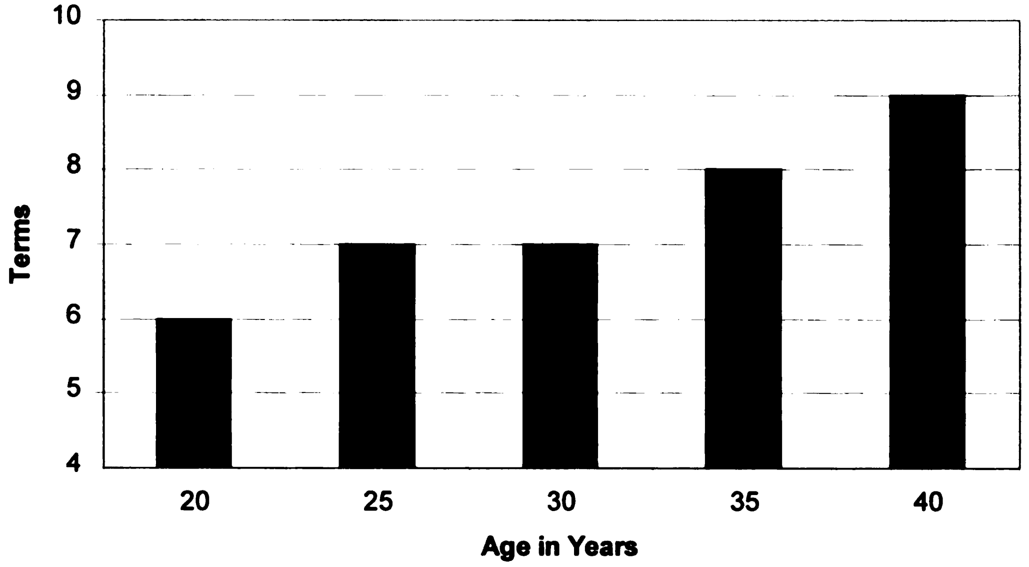
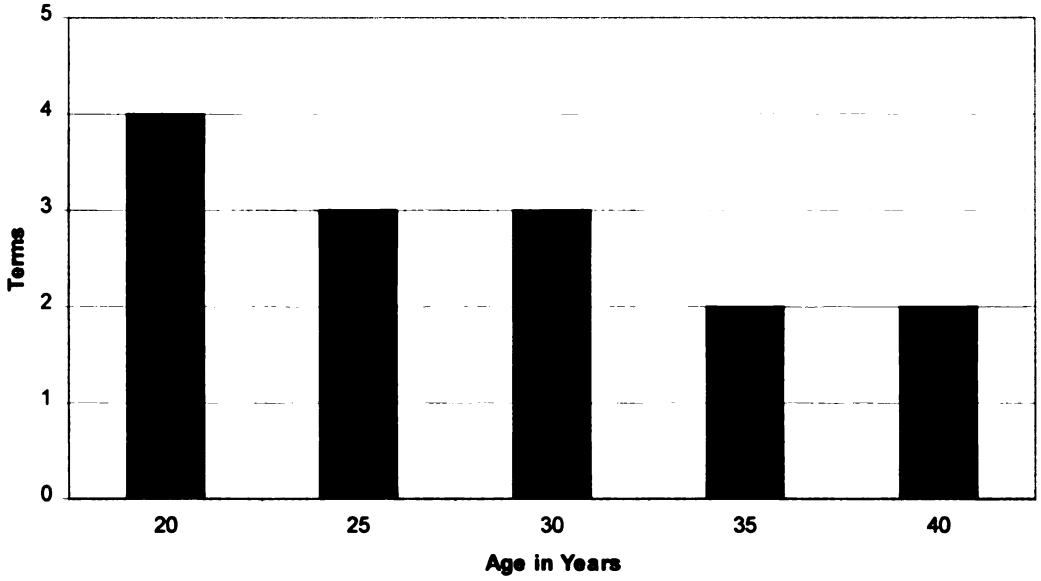


Figure 6b: Predicted Average Completion Time for Students Staying More Than One Term by Age Group on the Reading Comprehension Test



There are at least two apparent differences between students' completion rates on the listening and reading comprehension tests. First, older students fare much better on the reading comprehension test. If fact, the oldest group of students take the shortest amount of time (roughly 2 terms) to obtain a high enough logit score. The listening test has exactly the opposite findings. Older students take the longest time to obtain high enough logit scores. Second, the predicted time of completion between the listening and reading test greatly differ. Students on the reading test take a predicted maximum of 4 terms to obtain a high enough score. The listening test has a predicted maximum of 9 terms to obtain a sufficient score.

Clearly, students who score low on either exam have increasing growth rates. This suggests that the achievement use of the examination is justified. Older students are much more advantaged (i.e., perform better) on the reading test than the listening test.

Certain language groups also have significantly lower initial scores. On the listening test, Japanese, Chinese and Korean students have significantly lower initial scores. These lower initial scores will, of course, affect how long it would take to obtain sufficient scores to exit the program. On the reading test, Middle Eastern and Japanese students had significantly lower scores compared to other language groups. Again this will affect time necessary to obtain a sufficient score.

Knowledge of predicted exit times and specific demographic variables affecting initial status and/or growth is useful information for program administrators, university faculty (especially those with international student advisees), and student counselors. With the analysis provided here, it is possible to estimate the amount of time necessary to obtain a sufficient score on the listening and reading comprehension tests.

From a policy standpoint and for counseling purposes, the More variable is not very useful for predicting performance of incoming students. Figures 6a and 6b provide guidance on how long students are predicted to stay at the ELC once they have been at the Center for more than one term. However, none of the models presented thus far can be used directly for predicting incoming performance of students based upon investigated demographic variables. Table 45 presents fixed effects for listening and reading comprehension test models without the "More" variable. These results can be used to provide guidance on initial placement.

Table 45: Final Estimation of Demographic Variable Fixed Effects without the More Variable

```
Final estimation of fixed for effects Listening Test:
-------------------------------------------------------------------------
   Fixed Effect      Coefficient   Standard Error  T-ratio    P-value
-------------------------------------------------------------------------
 For        INTRCPT1, Π0
   INTRCPT2, B00       0.238996        0.124191         1.924     0.054
        AGE, B01      -0.022157        0.010264        -2.159     0.031
        UGR, B02       0.224570        0.094209         2.384     0.017
         ME, B03      -0.268531        0.184261        -1.457     0.145
        CHN, B04      -0.355692        0.171709        -2.071     0.038
        JPN, B05      -0.385474        0.151220        -2.549     0.011
        KRN, B06      -0.347397        0.135596        -2.562     0.011
 For   LINEAR slope, Π1
   INTRCPT2, B10       0.154433        0.020177         7.654     0.000
```

```
Final estimation of fixed effects for Reading Test:
-------------------------------------------------------------------------
   Fixed Effect      Coefficient   Standard Error  T-ratio    P-value
-------------------------------------------------------------------------
 For        INTRCPT1, Π0
   INTRCPT2, B00       1.069828        0.063839        16.758     0.000
         ME, B01      -0.689826        0.207913        -3.318     0.001
        JPN, B02      -0.400133        0.135800        -2.946     0.004
 For   LINEAR slope, Π1
   INTRCPT2, B10       0.553940        0.035218        15.729     0.000
```

The results reported above are similar to those found when the variable "More" was included in the model. For the listening model displayed in Table 45, age and language groups are significant predictors. That is, the older a student, the lower their initial listening score. Undergraduate students tend to have initial listening test scores that are 0.225 logits higher than their peers. Also, Chinese, Japanese and Korean students have significantly lower initial listening scores. Multivariate hypothesis tests indicate that these language groups do not differ from each other. Age seems not to be a factor for initial reading test scores. But, Middle Eastern and Japanese students have significantly lower initial reading test scores. The following table portrays predicted initial test scores and ELC completion times for incoming students based upon the estimates in Table 45.

Table 46 Predicted Initial Listening and Reading Comprehension Test Scores and Completion Time (in Terms) for Incoming Students

	Predicted Initial Logits		Predicted Terms to Completion	
Listening Test	20 year-olds	30 year-olds	20 year-olds	30 year-olds
Undergraduate	0.349	0.127	1	2
Other Group	0.124	-0.098	2	3
Middle Eastern	-0.145	-0.366	4	5
Chinese	-0.232	-0.453	4	5
Japanese	-0.261	-0.483	5	6
Korean	-0.223	-0.445	4	5

Reading Test	Predicted Initial Logits	Predicted Terms to Completion
Other Group	0.654	2
Middle Eastern	-0.035	3
Japanese	0.254	2

On the listening test, the undergraduate group has the highest predicted initial listening test logit and is predicted to have the shortest number of terms to completion. The Japanese language group has the lowest initial test logit and is predicted to take the longest to obtain sufficient test scores. To calculate term of completion, the average logit for students staying at the ELC for one term (from Table 44—0.367 for listening and 1.351 for reading) is used and compared with estimates obtained from Table 46 for initial status. For all language groups on the listening test, younger students are predicted to take the fewest terms to obtain sufficient test scores.

On the reading test, the Other language group is predicted to have the highest initial test logit and take the fewest terms to complete. Middle Eastern students are predicted to have the lowest initial reading test score and take the longest to obtain sufficient test scores to exit the ELC.

6.2 Study's Limitations

As with any study, several limitations impinge upon the generalizations possible with the findings. One obvious limitation is the sample size. Only 308 students had enough information for inclusion. Further, the number of terms (time-points) available for this study is greatly limited. Over 90% of the scores are collected within the first 3 terms. This makes powerful non-linear models difficult to investigate. The sample collected for this study is not atypical of the students attending American university English language programs. Because of MSU's admission requirements, many students in the ELC have TOEFL scores between 450 and 550. This is a relatively high level of English proficiency. Students with such scores would not be expected to stay in the ELC for extensive periods of time. What would the growth trajectories look like for students with very low levels of English proficiency on these examinations? At present this is unknown.

The person-level variables used for this study also limit generalizations. This study investigated how students' length of stay, language group, age, sex and academic status influence listening and reading comprehension test growth trajectories. Other equally important variables could effect test performance, e.g., personal motivation, personality, aptitude, preferred learning style, economic resources, teacher's ability, or instructed teaching methodology. Another limitation is the difficulty in ascertaining whether the observed growth trajectories are inherently a function of the test, the students, the students' language group, the students instructional program, or some combination of elements.

6.3 Study's Contribution and Future Directions

To date, no studies have reported on the investigation of a second or foreign language placement tests' growth trajectory. Research on second or foreign language proficiency or placement tests has looked predominantly at test validation in the most classical sense, i.e., reliability and construct or content validity (Davies, 1984; Wall, et al., 1994; and Fulcher, 1997). Wall, et al. (1997) and Fulcher (1997) have called for new ways of looking at second or foreign language test validity, especially as it relates to investigating gain scores or growth trajectories. This study provides an initial glimpse of growth trends for English as a second language listening and reading comprehension tests used for placement and achievement purposes at an American university. This study serves as a starting point in the investigation of growth on language tests. The study also provides a view into the differences between growth trajectories on listening and reading comprehension tests. This work introduces the notion of examining a test's growth trajectory to determine how different test uses might be appropriate or inappropriate.

The method of investigating change reported here might also prove useful in investigating curricular or classroom interventions. There is an emerging body of work investigating the decisions made about student achievement based upon gain (e.g., Millman, 1997). Many studies investigating interventions in the classroom, program, school, and district (new methodologies, teaching strategies, etc) are typically pretest, post-test design. The procedure used here may be used over more extended periods of time. This extension of time may better reflect long-term benefits as a function of specific interventions. Programmatically, this technique could also be used to identify characteristics of students who tend stay more than one term. This may assist counselors and admission officers in advising their students. This procedure may also be used as a mechanism for validating both placement and achievement test uses for university-based English as second language listening and reading comprehension tests.

APPENDIX A

Table 47: Classical and IRT Test Item Statistics for L92-1

Item	L92-1 Classical			L92-1 IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.74	0.24	0.22	-0.21	0.08	52.77	19
2	0.54	0.40	0.34	0.79	0.07	34.48	19
3	0.54	0.54	0.44	0.79	0.07	13.80	19
4	0.48	0.31	0.23	1.06	0.07	60.67	19
5	0.77	0.30	0.31	-0.41	0.09	18.36	19
6	0.73	0.47	0.44	-0.15	0.08	35.44	19
7	0.65	0.45	0.39	0.27	0.08	14.45	19
8	0.58	0.56	0.46	0.61	0.07	31.81	19
9	0.38	0.24	0.20	1.58	0.08	107.68	19
10	0.39	0.21	0.17	1.50	0.07	98.94	19
11	0.68	0.30	0.28	0.14	0.08	31.23	19
12	0.41	0.14	0.12	1.40	0.07	122.20	19
13	0.35	0.53	0.44	1.73	0.08	22.14	19
14	0.81	0.35	0.39	-0.69	0.09	24.84	19
15	0.53	0.54	0.43	0.84	0.07	14.14	19
16	0.73	0.46	0.43	-0.15	0.08	15.98	19
17	0.54	0.34	0.28	0.80	0.07	36.41	19
18	0.59	0.51	0.43	0.55	0.07	14.02	19
19	0.60	0.27	0.25	0.52	0.07	55.57	19
20	0.85	0.22	0.28	-1.00	0.10	13.02	19
21	0.54	0.56	0.44	0.80	0.07	17.04	19
22	0.81	0.15	0.16	-0.70	0.09	69.66	19
23	0.89	0.27	0.37	-1.34	0.11	15.19	19
24	0.96	0.11	0.27	-2.37	0.16	20.29	19
25	0.81	0.38	0.41	-0.64	0.09	26.83	19
26	0.69	0.51	0.46	0.09	0.08	20.87	19
27	0.81	0.38	0.43	-0.68	0.09	25.04	19
28	0.67	0.48	0.42	0.17	0.08	14.04	19
29	0.74	0.46	0.44	-0.20	0.08	17.30	19
30	0.42	0.47	0.37	1.35	0.07	22.44	19
31	0.25	0.41	0.38	2.27	0.08	22.37	19
32	0.55	0.57	0.44	0.76	0.07	24.04	19
33	0.64	0.48	0.44	0.31	0.08	21.06	19
34	0.69	0.66	0.60	0.05	0.08	88.28	19
35	0.90	0.20	0.32	-1.47	0.11	15.35	19
36	0.64	0.51	0.46	0.31	0.08	22.65	19
37	0.60	0.53	0.43	0.51	0.07	31.30	19
38	0.36	0.35	0.29	1.65	0.08	41.59	19
39	0.51	0.35	0.29	0.96	0.07	47.32	19
40	0.71	0.60	0.53	-0.06	0.08	47.99	19
41	0.95	0.10	0.22	-2.30	0.16	15.90	19
42	0.84	0.22	0.31	-0.89	0.10	13.98	19

Item	L92-1 Classical			L92-1 IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
43	0.92	0.18	0.27	-1.68	0.12	14.52	19
44	0.88	0.25	0.36	-1.23	0.11	13.30	19
45	0.82	0.31	0.36	-0.76	0.09	26.10	19
46	0.75	0.30	0.30	-0.27	0.08	21.91	19
47	0.95	0.13	0.30	-2.30	0.16	25.75	19
48	0.87	0.34	0.43	-1.17	0.10	34.87	19
49	0.77	0.32	0.31	-0.38	0.08	14.42	19
50	0.82	0.27	0.31	-0.78	0.09	28.05	19

Table 48: Classical and IRT Test Item Statistics for L92-2

Item	L92-2 Classical			L92-2 IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.54	0.29	0.24	0.05	0.08	26.45	19
2	0.48	0.08	0.09	0.35	0.08	109.40	19
3	0.80	0.27	0.28	-1.28	0.09	21.04	19
4	0.29	0.26	0.25	1.22	0.08	27.36	19
5	0.47	0.02	0.06	0.39	0.08	112.13	19
6	0.36	0.27	0.26	0.91	0.08	30.92	19
7	0.42	0.23	0.20	0.62	0.08	41.00	19
8	0.65	0.35	0.28	-0.44	0.08	23.43	19
9	0.34	0.14	0.16	0.97	0.08	60.87	19
10	0.43	0.23	0.22	0.54	0.08	58.51	19
11	0.76	0.20	0.18	-1.02	0.09	34.00	19
12	0.56	0.25	0.23	-0.02	0.08	30.58	19
13	0.54	0.45	0.34	0.08	0.08	27.01	19
14	0.81	0.31	0.30	-1.36	0.09	30.03	19
15	0.48	0.05	0.01	0.34	0.08	166.70	19
16	0.64	0.46	0.39	-0.39	0.08	18.69	19
17	0.71	0.53	0.47	-0.72	0.08	51.54	19
18	0.52	0.28	0.26	0.17	0.08	24.23	19
19	0.35	0.10	0.11	0.94	0.08	66.22	19
20	0.31	0.18	0.18	1.13	0.08	49.77	19
21	0.18	0.19	0.19	1.94	0.10	31.42	19
22	0.68	0.44	0.38	-0.57	0.08	30.05	19
23	0.70	0.39	0.34	-0.71	0.08	30.74	19
24	0.73	0.38	0.34	-0.83	0.08	47.56	19
25	0.21	0.16	0.19	1.75	0.09	35.62	19
26	0.71	0.54	0.46	-0.72	0.08	36.99	19
27	0.33	0.40	0.34	1.03	0.08	20.92	19
28	0.38	0.42	0.36	0.78	0.08	14.98	19
29	0.45	0.42	0.34	0.45	0.08	10.69	19
30	0.77	0.36	0.37	-1.07	0.09	39.70	19
31	0.97	0.03	0.10	-3.37	0.20	16.27	19
32	0.48	0.22	0.23	0.33	0.08	52.04	19

Item	L92-2 Classical			L92-2 IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
33	0.81	0.40	0.41	-1.35	0.09	37.68	19
34	0.41	0.21	0.18	0.66	0.08	63.20	19
35	0.53	0.48	0.38	0.13	0.08	13.96	19
36	0.23	0.25	0.28	1.56	0.09	18.83	19
37	0.64	0.52	0.40	-0.38	0.08	32.79	19
38	0.78	0.40	0.36	-1.12	0.09	28.10	19
39	0.57	0.47	0.39	-0.08	0.08	16.47	19
40	0.77	0.33	0.31	-1.09	0.09	27.89	19
41	0.80	0.26	0.30	-1.29	0.09	18.47	19
42	0.45	0.63	0.52	0.46	0.08	80.03	19
43	0.46	0.69	0.56	0.44	0.08	89.27	19
44	0.42	0.64	0.52	0.58	0.08	67.01	19
45	0.90	0.21	0.28	-2.10	0.12	26.14	19
46	0.48	0.64	0.52	0.33	0.08	90.86	19
47	0.47	0.60	0.51	0.38	0.08	62.11	19
48	0.52	0.58	0.47	0.16	0.08	53.90	19
49	0.43	0.64	0.54	0.54	0.08	80.94	19
50	0.83	0.32	0.34	-1.48	0.10	29.11	19
51	0.47	0.68	0.55	0.38	0.08	82.64	19
52	0.44	0.64	0.54	0.52	0.08	68.35	19
53	0.29	0.32	0.33	1.25	0.08	56.38	19

Table 49: Classical and IRT Test Item Statistics for R91-1a

Item	R91-1a Classical			R91-1a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.64	0.88	0.75	-0.61	0.13	48.07	19
2	0.88	0.21	0.28	-1.87	0.18	26.45	19
3	0.63	0.87	0.71	-0.58	0.13	47.42	19
4	0.65	0.70	0.59	-0.66	0.13	9.96	19
5	0.48	0.79	0.62	0.06	0.13	27.75	19
6	0.60	0.32	0.26	-0.43	0.13	81.97	19
7	0.63	0.18	0.16	-0.55	0.13	138.83	19
8	0.37	0.60	0.51	0.55	0.13	20.08	19
9	0.39	0.60	0.49	0.45	0.13	24.34	19
10	0.28	0.46	0.41	0.95	0.14	19.58	19
11	0.40	0.68	0.57	0.41	0.13	23.53	19
12	0.52	0.75	0.59	-0.11	0.13	28.65	19
13	0.65	0.90	0.75	-0.66	0.13	65.65	19
14	0.61	0.70	0.58	-0.49	0.13	23.68	19
15	0.68	0.88	0.77	-0.77	0.14	62.13	19
16	0.63	0.61	0.52	-0.56	0.13	27.97	19
17	0.63	0.78	0.67	-0.56	0.13	29.33	19
18	0.44	0.73	0.59	0.25	0.13	27.80	19
19	0.76	0.26	0.28	-1.19	0.15	53.95	19

Item	R91-1a Classical			R91-1a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
20	0.33	0.52	0.42	0.71	0.14	36.83	19
21	0.52	0.57	0.48	-0.10	0.13	27.79	19
22	0.28	0.38	0.35	0.95	0.14	46.52	19
23	0.54	0.89	0.73	-0.18	0.13	29.96	19
24	0.70	0.49	0.47	-0.87	0.14	18.55	19
25	0.62	0.32	0.32	-0.51	0.13	53.72	19
26	0.50	0.65	0.52	-0.02	0.13	7.31	19
27	0.55	0.06	0.04	-0.23	0.13	216.01	19
28	0.40	0.59	0.49	0.39	0.13	19.66	19
29	0.68	0.42	0.43	-0.80	0.14	52.09	19
30	0.44	0.89	0.74	0.22	0.13	45.83	19
31	0.56	0.88	0.76	-0.25	0.13	37.63	19
32	0.60	0.89	0.75	-0.44	0.13	41.60	19
33	0.40	0.47	0.37	0.42	0.13	77.32	19
34	0.52	-0.17	-0.12	-0.11	0.13	295.76	19
35	0.52	0.85	0.68	-0.11	0.13	42.35	19
36	0.44	0.62	0.52	0.25	0.13	22.83	19
37	0.56	0.88	0.74	-0.26	0.13	40.04	19
38	0.50	0.77	0.67	-0.04	0.13	25.37	19
39	0.34	0.56	0.50	0.67	0.14	41.81	19
40	0.42	0.64	0.52	0.32	0.13	24.91	19
41	0.77	0.51	0.55	-1.20	0.15	24.72	19
42	0.39	0.65	0.56	0.43	0.13	19.87	19
43	0.50	0.84	0.69	-0.01	0.13	29.95	19
44	0.46	0.73	0.57	0.15	0.13	19.31	19
45	0.50	0.78	0.65	-0.04	0.13	26.82	19
46	0.46	0.76	0.64	0.17	0.13	21.09	19
47	0.50	0.75	0.61	0.00	0.13	11.74	19
48	0.58	0.56	0.44	-0.34	0.13	25.37	19
49	0.61	0.83	0.74	-0.49	0.13	38.68	19
50	0.48	0.05	0.06	0.08	0.13	181.41	19
51	0.75	0.25	0.29	-1.10	0.14	81.11	19

Table 50: Classical and IRT Test Item Statistics for R91-1b

Item	R91-1b Classical			R91-1b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.89	0.30	0.35	-1.39	0.28	14.31	38
2	0.91	0.20	0.30	-1.74	0.31	8.94	38
3	0.91	0.30	0.39	-1.64	0.30	9.09	38
4	0.77	0.50	0.48	-0.41	0.22	19.63	38
5	0.86	0.33	0.41	-1.11	0.26	13.75	38
6	0.52	0.28	0.27	0.89	0.19	32.86	38
7	0.60	0.58	0.47	0.54	0.19	15.75	38
8	0.95	0.07	0.18	-2.34	0.39	34.13	38

Item	R91-1b Classical			R91-1b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
9	0.81	0.21	0.22	-0.70	0.23	21.68	38
10	0.95	0.08	0.20	-2.34	0.39	33.24	38
11	0.81	0.28	0.33	-0.70	0.23	17.67	38
12	0.84	0.18	0.26	-0.98	0.25	19.17	38
13	0.55	0.44	0.38	0.75	0.19	16.58	38
14	0.72	0.48	0.45	-0.10	0.21	10.71	38
15	0.41	0.23	0.18	1.44	0.19	32.55	38
16	0.59	0.34	0.30	0.58	0.19	24.57	38
17	0.35	0.09	0.06	1.73	0.19	42.09	38
18	0.74	0.50	0.54	-0.27	0.21	13.09	38
19	0.82	0.33	0.38	-0.81	0.24	17.13	38
20	0.75	0.60	0.55	-0.32	0.21	19.04	38
21	0.65	0.24	0.28	0.29	0.20	20.51	38
22	0.67	0.42	0.35	0.18	0.20	19.42	38
23	0.60	0.39	0.33	0.54	0.19	20.92	38
24	0.33	0.41	0.35	1.88	0.20	12.81	38
25	0.52	0.64	0.47	0.93	0.19	19.32	38
26	0.68	0.51	0.42	0.10	0.20	18.32	38
27	0.75	0.45	0.37	-0.32	0.21	13.66	38
28	0.50	0.42	0.34	1.03	0.19	18.39	38
29	0.50	0.59	0.49	1.00	0.19	13.67	38
30	0.92	0.15	0.24	-1.84	0.32	19.32	38
31	0.57	0.49	0.40	0.68	0.19	16.03	38
32	0.70	0.63	0.56	0.02	0.20	9.11	38
33	0.62	0.34	0.31	0.44	0.19	29.29	38
34	0.72	0.53	0.48	-0.10	0.21	10.75	38
35	0.61	0.59	0.51	0.47	0.19	16.19	38
36	0.67	0.68	0.56	0.14	0.20	15.05	38
37	0.67	0.70	0.64	0.14	0.20	16.05	38
38	0.87	0.38	0.53	-1.17	0.26	13.03	38
39	0.44	0.25	0.25	1.30	0.19	20.17	38
40	0.77	0.56	0.56	-0.45	0.22	18.01	38
41	0.48	0.57	0.43	1.10	0.19	19.77	38
42	0.49	0.47	0.39	1.06	0.19	20.33	38
43	0.79	0.55	0.58	-0.55	0.22	17.50	38
44	0.66	0.56	0.51	0.21	0.20	12.46	38
45	0.80	0.50	0.57	-0.65	0.23	12.99	38
46	0.79	0.46	0.52	-0.60	0.23	13.39	38
47	0.48	0.44	0.37	1.13	0.19	19.97	38
48	0.35	0.29	0.19	1.77	0.19	29.51	38
49	0.70	0.58	0.51	-0.02	0.20	12.82	38
50	0.57	0.56	0.44	0.65	0.19	16.90	38
51	0.77	0.45	0.51	-0.45	0.22	16.13	38

Table 51: Classical and IRT Test Item Statistics for R91-2a

Item	R91-2a Classical			R91-2a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.79	0.36	0.42	-0.38	0.15	25.22	19
2	0.74	0.25	0.26	-0.06	0.14	21.24	19
3	0.76	0.31	0.34	-0.21	0.14	15.20	19
4	0.89	0.27	0.44	-1.24	0.19	11.43	19
5	0.76	0.45	0.51	-0.21	0.14	18.86	19
6	0.71	0.38	0.39	0.09	0.13	17.85	19
7	0.65	0.43	0.43	0.43	0.13	20.33	19
8	0.88	0.32	0.48	-1.11	0.18	16.66	19
9	0.41	0.29	0.18	1.61	0.12	48.69	19
10	0.65	0.31	0.26	0.45	0.13	21.61	19
11	0.68	0.32	0.32	0.28	0.13	15.68	19
12	0.41	0.35	0.27	1.59	0.12	34.18	19
13	0.78	0.44	0.51	-0.32	0.15	15.24	19
14	0.79	0.19	0.19	-0.42	0.15	43.42	19
15	0.61	0.40	0.32	0.66	0.13	13.33	19
16	0.40	0.22	0.18	1.67	0.12	41.89	19
17	0.81	0.26	0.36	-0.51	0.15	19.36	19
18	0.69	0.47	0.36	0.23	0.13	24.76	19
19	0.49	0.45	0.37	1.22	0.12	19.14	19
20	0.65	0.37	0.35	0.45	0.13	20.16	19
21	0.67	0.34	0.29	0.32	0.13	34.14	19
22	0.89	0.29	0.45	-1.31	0.19	13.25	19
23	0.93	0.16	0.36	-1.83	0.23	28.31	19
24	0.91	0.28	0.52	-1.55	0.21	24.78	19
25	0.44	0.56	0.45	1.46	0.12	8.81	19
26	0.61	0.45	0.39	0.66	0.13	10.47	19
27	0.85	0.38	0.49	-0.87	0.17	15.23	19
28	0.80	0.44	0.52	-0.49	0.15	25.42	19
29	0.86	0.31	0.45	-0.92	0.17	12.70	19
30	0.90	0.23	0.37	-1.42	0.20	6.74	19
31	0.60	0.46	0.38	0.70	0.12	14.45	19
32	0.41	0.14	0.13	1.59	0.12	50.61	19
33	0.76	0.54	0.54	-0.21	0.14	24.74	19
34	0.57	0.53	0.43	0.84	0.12	15.86	19
35	0.66	0.45	0.39	0.37	0.13	19.85	19
36	0.52	0.51	0.42	1.05	0.12	32.96	19
37	0.69	0.39	0.34	0.20	0.13	20.75	19
38	0.69	0.40	0.37	0.22	0.13	15.44	19
39	0.70	0.58	0.54	0.18	0.13	26.40	19
40	0.89	0.23	0.34	-1.27	0.19	12.88	19
41	0.70	0.50	0.44	0.18	0.13	15.26	19
42	0.84	0.38	0.45	-0.79	0.16	16.88	19
43	0.68	0.49	0.40	0.27	0.13	27.82	19

Item	R91-2a Classical			R91-2a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
44	0.72	0.37	0.37	0.04	0.14	18.51	19
45	0.75	0.44	0.44	-0.11	0.14	21.39	19
46	0.90	0.22	0.39	-1.42	0.20	13.81	19
47	0.91	0.12	0.20	-1.51	0.20	22.84	19
48	0.77	0.49	0.42	-0.27	0.14	22.16	19
49	0.39	0.32	0.24	1.69	0.13	31.73	19

Table 52: Classical and IRT Test Item Statistics for R91-2b

Item	R91-2b Classical			R91-2b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.80	0.26	0.32	-0.45	0.16	12.74	19
2	0.79	0.33	0.36	-0.35	0.15	23.46	19
3	0.76	0.43	0.43	-0.19	0.15	18.26	19
4	0.77	0.35	0.38	-0.24	0.15	28.68	19
5	0.88	0.21	0.37	-1.13	0.19	16.39	19
6	0.63	0.14	0.17	0.54	0.13	47.27	19
7	0.67	0.23	0.22	0.33	0.14	28.22	19
8	0.40	0.33	0.28	1.69	0.13	12.09	19
9	0.76	0.51	0.55	-0.15	0.15	19.36	19
10	0.47	0.40	0.31	1.33	0.13	23.90	19
11	0.67	0.29	0.31	0.33	0.14	12.35	19
12	0.64	0.30	0.26	0.49	0.13	16.32	19
13	0.87	0.28	0.49	-1.06	0.19	19.62	19
14	0.94	0.13	0.29	-1.91	0.25	13.92	19
15	0.95	0.14	0.43	-2.19	0.28	13.96	19
16	0.78	0.21	0.26	-0.30	0.15	32.05	19
17	0.63	0.41	0.36	0.56	0.13	26.71	19
18	0.45	0.28	0.21	1.45	0.13	36.38	19
19	0.83	0.32	0.39	-0.68	0.17	19.76	19
20	0.69	0.49	0.43	0.22	0.14	14.08	19
21	0.77	0.26	0.27	-0.24	0.15	19.18	19
22	0.66	0.40	0.34	0.38	0.13	11.86	19
23	0.91	0.23	0.52	-1.45	0.21	17.12	19
24	0.47	0.31	0.27	1.33	0.13	21.71	19
25	0.73	0.37	0.31	0.00	0.14	26.38	19
26	0.70	0.38	0.40	0.20	0.14	18.22	19
27	0.78	0.46	0.51	-0.28	0.15	24.56	19
28	0.89	0.25	0.46	-1.24	0.20	18.67	19
29	0.93	0.07	0.24	-1.85	0.24	20.84	19
30	0.77	0.43	0.52	-0.24	0.15	23.83	19
31	0.37	0.25	0.20	1.81	0.13	50.03	19
32	0.64	0.50	0.45	0.53	0.13	15.33	19
33	0.54	0.49	0.42	1.01	0.13	14.52	19
34	0.73	0.41	0.45	0.02	0.14	17.51	19

Item	R91-2b Classical			R91-2b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
35	0.59	0.55	0.44	0.76	0.13	22.75	19
36	0.72	0.60	0.58	0.06	0.14	25.23	19
37	0.81	0.30	0.42	-0.52	0.16	22.22	19
38	0.69	0.51	0.49	0.26	0.14	27.22	19
39	0.81	0.36	0.51	-0.50	0.16	21.24	19
40	0.50	0.58	0.41	1.20	0.13	19.16	19
41	0.60	0.44	0.41	0.70	0.13	15.00	19
42	0.81	0.35	0.43	-0.55	0.16	12.41	19
43	0.69	0.43	0.43	0.26	0.14	12.93	19
44	0.88	0.30	0.50	-1.13	0.19	13.18	19
45	0.91	0.19	0.44	-1.54	0.22	9.55	19
46	0.51	0.52	0.42	1.17	0.13	20.75	19
47	0.40	0.05	0.07	1.69	0.13	77.97	19
48	0.83	0.43	0.54	-0.66	0.17	26.37	19
49	0.64	0.51	0.42	0.49	0.13	18.95	19

Table 53: Classical and IRT Test Item Statistics for R91-3a

Item	R91-3a Classical			R91-3a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.63	0.35	0.28	0.25	0.16	27.65	19
2	0.82	0.40	0.47	-0.86	0.20	17.75	19
3	0.75	0.36	0.31	-0.38	0.18	13.32	19
4	0.88	0.18	0.25	-1.39	0.23	13.81	19
5	0.80	0.27	0.30	-0.68	0.19	14.83	19
6	0.84	0.26	0.22	-1.02	0.21	29.66	19
7	0.65	0.57	0.48	0.17	0.16	14.05	19
8	0.66	0.17	0.19	0.09	0.16	25.78	19
9	0.64	0.37	0.33	0.22	0.16	18.06	19
10	0.81	0.36	0.32	-0.78	0.19	27.95	19
11	0.49	0.35	0.26	0.92	0.16	26.57	19
12	0.50	0.48	0.40	0.90	0.16	14.94	19
13	0.69	0.33	0.28	-0.01	0.17	11.36	19
14	0.49	0.28	0.24	0.92	0.16	20.33	19
15	0.37	0.26	0.23	1.51	0.16	47.40	19
16	0.54	0.30	0.26	0.68	0.16	19.36	19
17	0.95	0.13	0.28	-2.44	0.34	22.50	19
18	0.94	0.16	0.33	-2.23	0.31	9.95	19
19	0.96	0.05	0.11	-2.71	0.38	22.37	19
20	0.77	0.36	0.33	-0.48	0.18	22.10	19
21	0.57	0.40	0.33	0.54	0.16	19.28	19
22	0.72	0.44	0.34	-0.21	0.17	23.75	19
23	0.72	0.51	0.54	-0.21	0.17	27.88	19
24	0.40	0.39	0.34	1.36	0.16	25.05	19
25	0.88	0.27	0.35	-1.34	0.23	17.87	19

Item	R91-3a Classical			R91-3a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
26	0.20	0.36	0.36	2.48	0.19	21.36	19
27	0.49	0.40	0.33	0.92	0.16	17.15	19
28	0.52	0.26	0.18	0.80	0.16	36.76	19
29	0.89	0.22	0.31	-1.50	0.24	11.31	19
30	0.76	0.38	0.34	-0.45	0.18	13.60	19
31	0.62	0.46	0.39	0.30	0.16	26.89	19
32	0.70	0.64	0.56	-0.07	0.17	22.87	19
33	0.76	0.56	0.55	-0.42	0.18	22.88	19
34	0.57	0.62	0.50	0.54	0.16	20.78	19
35	0.82	0.31	0.38	-0.82	0.19	12.66	19
36	0.88	0.31	0.43	-1.34	0.23	19.19	19
37	0.65	0.58	0.51	0.17	0.16	14.46	19
38	0.46	0.46	0.39	1.06	0.16	12.71	19
39	0.64	0.51	0.43	0.19	0.16	19.13	19
40	0.67	0.47	0.41	0.07	0.16	13.49	19
41	0.54	0.59	0.49	0.71	0.16	21.15	19
42	0.49	0.37	0.31	0.95	0.16	22.79	19
43	0.44	0.59	0.43	1.18	0.16	17.75	19
44	0.76	0.51	0.50	-0.42	0.18	19.95	19
45	0.70	0.42	0.37	-0.07	0.17	19.30	19
46	0.46	0.50	0.40	1.09	0.16	22.30	19
47	0.37	0.06	0.06	1.53	0.16	67.77	19
48	0.86	0.29	0.35	-1.14	0.21	18.90	19
49	0.86	0.20	0.32	-1.19	0.22	23.92	19
50	0.69	0.67	0.55	-0.04	0.17	32.46	19
51	0.48	0.44	0.35	0.97	0.16	12.24	19
52	0.49	0.60	0.47	0.92	0.16	10.99	19
53	0.53	0.48	0.44	0.73	0.16	20.64	19

Table 54: Classical and IRT Test Item Statistics for R91-3b

Item	R91-3b Classical			R91-3b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.71	0.20	0.16	-0.21	0.18	25.49	19
2	0.89	0.22	0.32	-1.52	0.25	13.08	19
3	0.76	0.29	0.26	-0.51	0.19	31.78	19
4	0.72	0.43	0.42	-0.24	0.18	13.41	19
5	0.81	0.40	0.47	-0.82	0.20	18.09	19
6	0.86	0.28	0.43	-1.23	0.23	18.08	19
7	0.51	0.40	0.29	0.80	0.16	25.03	19
8	0.74	0.37	0.39	-0.34	0.18	23.68	19
9	0.79	0.45	0.48	-0.66	0.20	18.16	19
10	0.34	0.27	0.24	1.59	0.17	23.69	19
11	0.92	0.23	0.42	-1.87	0.28	10.70	19
12	0.23	0.27	0.20	2.21	0.19	34.53	19

Item	R91-3b Classical			R91-3b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
13	0.59	0.30	0.29	0.43	0.17	19.69	19
14	0.55	0.30	0.29	0.59	0.17	27.49	19
15	0.60	0.54	0.48	0.37	0.17	19.66	19
16	0.41	0.39	0.33	1.26	0.17	23.46	19
17	0.40	0.35	0.33	1.31	0.17	17.04	19
18	0.60	0.46	0.36	0.35	0.17	17.33	19
19	0.64	0.60	0.48	0.18	0.17	20.50	19
20	0.60	0.38	0.35	0.37	0.17	18.47	19
21	0.64	0.29	0.25	0.18	0.17	24.11	19
22	0.82	0.43	0.47	-0.86	0.21	15.13	19
23	0.45	0.34	0.26	1.04	0.17	28.99	19
24	0.47	0.32	0.26	0.99	0.16	15.26	19
25	0.95	0.17	0.35	-2.36	0.34	11.48	19
26	0.90	0.19	0.28	-1.65	0.26	13.59	19
27	0.94	0.13	0.20	-2.15	0.32	20.55	19
28	0.75	0.31	0.33	-0.44	0.19	20.22	19
29	0.78	0.51	0.55	-0.62	0.20	22.80	19
30	0.71	0.39	0.38	-0.18	0.18	23.75	19
31	0.47	0.48	0.41	0.99	0.16	17.14	19
32	0.41	-0.01	0.02	1.26	0.17	58.08	19
33	0.83	0.43	0.55	-0.95	0.21	23.28	19
34	0.86	0.34	0.48	-1.23	0.23	15.00	19
35	0.72	0.60	0.55	-0.27	0.18	20.64	19
36	0.48	0.16	0.16	0.91	0.16	28.45	19
37	0.43	0.71	0.54	1.17	0.17	24.98	19
38	0.58	0.15	0.19	0.46	0.17	39.44	19
39	0.88	0.26	0.38	-1.40	0.24	10.07	19
40	0.70	0.24	0.23	-0.15	0.18	18.11	19
41	0.66	0.50	0.47	0.06	0.17	19.19	19
42	0.73	0.39	0.39	-0.31	0.18	14.18	19
43	0.75	0.45	0.48	-0.41	0.19	19.14	19
44	0.52	0.54	0.44	0.72	0.16	25.46	19
45	0.75	0.34	0.37	-0.44	0.19	13.47	19
46	0.84	0.38	0.53	-1.08	0.22	19.56	19
47	0.70	0.41	0.32	-0.12	0.18	20.17	19
48	0.39	0.42	0.36	1.34	0.17	24.55	19
49	0.63	0.45	0.34	0.24	0.17	17.63	19
50	0.67	0.35	0.32	0.03	0.17	13.31	19
51	0.40	0.55	0.41	1.31	0.17	18.93	19
52	0.45	0.32	0.27	1.04	0.17	15.61	19
53	0.50	0.42	0.33	0.83	0.16	21.95	19

Table 55: Classical and IRT Test Item Statistics for R92-1a

Item	R92-1a Classical			R92-1a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.69	0.59	0.55	-0.07	0.18	16.94	19
2	0.75	0.40	0.42	-0.42	0.19	14.91	19
3	0.87	0.24	0.33	-1.37	0.23	16.85	19
4	0.83	0.31	0.39	-1.03	0.21	17.35	19
5	0.70	0.38	0.39	-0.16	0.18	8.92	19
6	0.68	0.14	0.16	-0.01	0.17	41.84	19
7	0.57	0.65	0.48	0.55	0.17	11.87	19
8	0.50	0.45	0.35	0.91	0.16	35.34	19
9	0.85	0.41	0.48	-1.17	0.22	12.38	19
10	0.62	0.59	0.51	0.28	0.17	21.94	19
11	0.53	0.63	0.44	0.76	0.16	42.25	19
12	0.54	0.50	0.46	0.68	0.16	13.26	19
13	0.72	0.49	0.45	-0.25	0.18	11.40	19
14	0.72	0.62	0.53	-0.29	0.18	17.48	19
15	0.74	0.58	0.53	-0.38	0.18	18.28	19
16	0.71	0.66	0.57	-0.19	0.18	14.67	19
17	0.71	0.51	0.46	-0.22	0.18	9.91	19
18	0.66	0.45	0.39	0.11	0.17	18.55	19
19	0.73	0.53	0.53	-0.32	0.18	14.98	19
20	0.52	0.16	0.16	0.81	0.16	44.71	19
21	0.55	0.47	0.39	0.65	0.16	23.08	19
22	0.84	0.33	0.40	-1.07	0.21	8.35	19
23	0.58	0.35	0.30	0.50	0.17	33.89	19
24	0.74	0.55	0.52	-0.35	0.18	15.67	19
25	0.34	0.19	0.15	1.71	0.17	54.84	19
26	0.89	0.27	0.42	-1.59	0.25	18.08	19
27	0.78	0.47	0.50	-0.67	0.19	17.25	19
28	0.85	0.25	0.35	-1.22	0.22	11.61	19
29	0.73	0.62	0.56	-0.32	0.18	17.47	19
30	0.35	0.35	0.31	1.62	0.17	15.69	19
31	0.83	0.37	0.48	-1.03	0.21	16.82	19
32	0.72	0.55	0.54	-0.29	0.18	13.33	19
33	0.83	0.52	0.60	-0.99	0.21	20.58	19
34	0.53	0.60	0.48	0.76	0.16	13.99	19
35	0.77	0.60	0.63	-0.59	0.19	21.17	19
36	0.34	0.22	0.16	1.68	0.17	78.92	19
37	0.56	0.68	0.55	0.60	0.16	19.96	19
38	0.66	0.58	0.48	0.08	0.17	15.59	19
39	0.78	0.58	0.58	-0.67	0.19	13.83	19
40	0.49	0.39	0.29	0.94	0.16	25.89	19
41	0.42	0.53	0.42	1.30	0.16	21.10	19
42	0.49	0.43	0.37	0.97	0.16	29.49	19
43	0.74	0.54	0.53	-0.38	0.18	11.08	19

Item	R92-1a Classical			R92-1a IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
44	0.70	0.43	0.40	-0.13	0.18	22.98	19
45	0.86	0.37	0.52	-1.26	0.22	50.31	19
46	0.87	0.26	0.41	-1.42	0.23	16.18	19
47	0.53	0.37	0.30	0.76	0.16	25.01	19
48	0.34	0.32	0.27	1.68	0.17	21.54	19
49	0.75	0.54	0.48	-0.42	0.19	16.72	19
50	0.50	0.61	0.45	0.91	0.16	26.53	19

Table 56: Classical and IRT Test Item Statistics for R92-1b

Item	R92-1b Classical			R92-1b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
1	0.74	0.48	0.45	-0.53	0.18	16.36	19
2	0.67	0.55	0.48	-0.15	0.17	19.88	19
3	0.84	0.37	0.37	-1.26	0.21	17.60	19
4	0.80	0.50	0.56	-0.96	0.20	19.10	19
5	0.58	0.78	0.59	0.37	0.17	20.13	19
6	0.53	0.54	0.38	0.61	0.16	17.86	19
7	0.54	0.33	0.23	0.58	0.16	47.83	19
8	0.63	0.61	0.51	0.09	0.17	7.73	19
9	0.71	0.47	0.38	-0.33	0.18	14.65	19
10	0.71	0.57	0.51	-0.36	0.18	24.18	19
11	0.75	0.51	0.47	-0.59	0.19	14.00	19
12	0.73	0.36	0.39	-0.49	0.18	19.32	19
13	0.61	0.48	0.41	0.23	0.17	17.03	19
14	0.73	0.66	0.56	-0.46	0.18	21.31	19
15	0.49	0.47	0.36	0.82	0.16	29.12	19
16	0.62	0.59	0.49	0.17	0.17	21.04	19
17	0.65	0.47	0.42	0.00	0.17	15.46	19
18	0.65	0.29	0.29	0.00	0.17	43.91	19
19	0.49	0.62	0.45	0.79	0.16	13.06	19
20	0.48	0.44	0.35	0.87	0.16	27.26	19
21	0.82	0.35	0.40	-1.13	0.21	13.36	19
22	0.49	0.39	0.31	0.79	0.16	32.68	19
23	0.75	0.47	0.43	-0.59	0.19	24.71	19
24	0.83	0.32	0.42	-1.17	0.21	19.33	19
25	0.29	0.23	0.18	1.88	0.18	47.13	19
26	0.43	0.36	0.28	1.11	0.16	26.84	19
27	0.46	0.41	0.34	0.95	0.16	26.44	19
28	0.72	0.66	0.62	-0.40	0.18	19.87	19
29	0.66	0.54	0.45	-0.09	0.17	17.25	19
30	0.84	0.52	0.63	-1.30	0.22	26.18	19
31	0.84	0.39	0.50	-1.30	0.22	23.78	19
32	0.43	0.38	0.32	1.14	0.16	25.16	19
33	0.39	0.22	0.22	1.33	0.17	38.21	19

Item	R92-1b Classical			R92-1b IRT			
	P-Value	Disc.	Pt. Bis.	Diff	SE	Chi-sq	df
34	0.70	0.75	0.68	-0.30	0.18	27.92	19
35	0.51	0.57	0.46	0.74	0.16	15.26	19
36	0.88	0.33	0.45	-1.72	0.24	17.63	19
37	0.71	0.66	0.60	-0.33	0.18	16.12	19
38	0.79	0.49	0.54	-0.88	0.20	20.18	19
39	0.70	0.66	0.60	-0.30	0.18	19.11	19
40	0.33	0.35	0.28	1.67	0.17	27.96	19
41	0.80	0.37	0.44	-0.96	0.20	15.92	19
42	0.76	0.40	0.39	-0.70	0.19	9.12	19
43	0.78	0.54	0.61	-0.85	0.19	25.65	19
44	0.51	0.59	0.49	0.74	0.16	8.45	19
45	0.70	0.69	0.68	-0.30	0.18	39.70	19
46	0.37	0.20	0.17	1.41	0.17	55.40	19
47	0.45	0.44	0.39	1.01	0.16	17.92	19
48	0.67	0.49	0.44	-0.15	0.17	15.99	19
49	0.70	0.66	0.58	-0.30	0.18	14.81	19
50	0.54	0.33	0.28	0.58	0.16	24.25	19

APPENDIX B

Figure 7: L92-1 Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One					
Factor	1	2	3	4	5
Eigenvalue	11.82	2.41	1.96	1.66	1.58
Factor	6	7	8	9	10
Eigenvalue	1.53	1.45	1.42	1.32	1.30
Factor	11	12	13	14	15
Eigenvalue	1.22	1.17	1.11	1.06	1.04
Factor	16				
Eigenvalue	1.00				

Percent of Total Variance Explained					
Factor	1	2	3	4	5
Percent Variance	23.65	4.83	3.92	3.31	3.16
Factor	6	7	8	9	10
Percent Variance	3.06	2.90	2.83	2.65	2.60
Factor	11	12	13	14	15
Percent Variance	2.43	2.34	2.23	2.11	2.07
Factor	16				
Percent Variance	2.01				

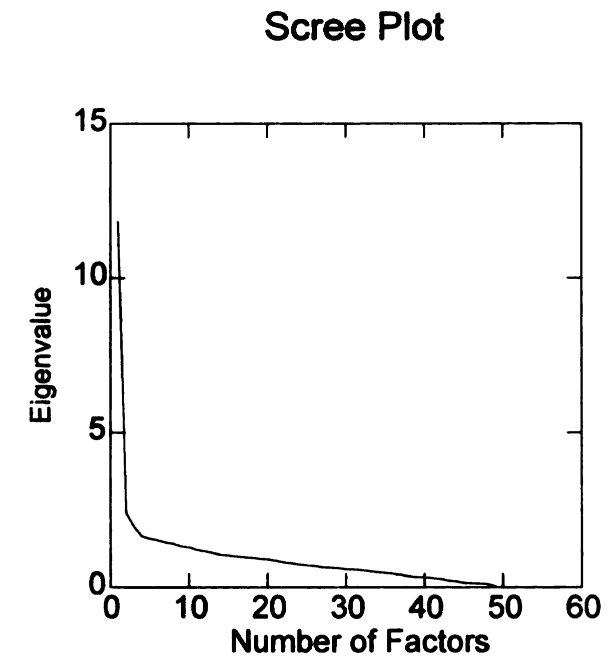


Figure 8: L92-2 Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	10.10	5.06	2.79	2.10	1.72
Factor	6	7	8	9	10
Eigenvalue	1.66	1.52	1.34	1.29	1.28
Factor	11	12	13	14	15
Eigenvalue	1.24	1.19	1.17	1.10	1.07
Factor	16				
Eigenvalue	1.05				

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	19.06	9.54	5.27	3.97	3.24
Factor	6	7	8	9	10
Percent Variance	3.13	2.87	2.52	2.43	2.41
Factor	11	12	13	14	15
Percent Variance	2.33	2.24	2.20	2.07	2.03
Factor	16				
Percent Variance	1.97				

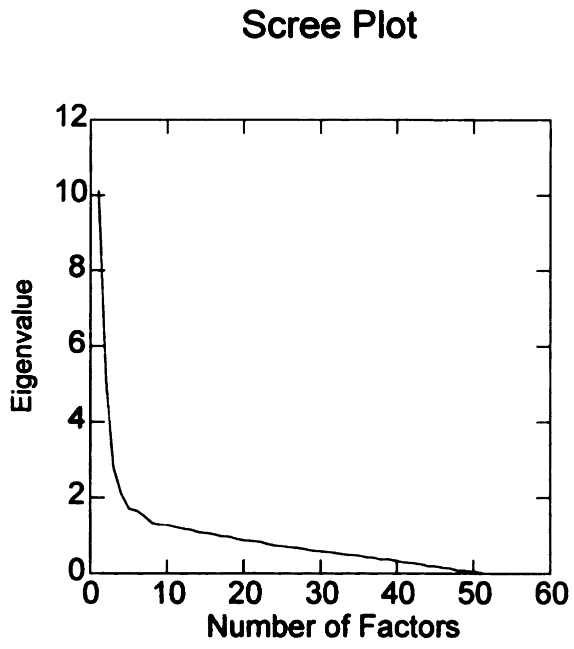


Figure 9: R91-1a Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	23.559	3.934	2.333	1.512	1.295
Factor	6	7	8	9	10
Eigenvalue	1.257	1.19	1.136	1.108	1.048

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	46.194	7.714	4.575	2.966	2.539
Factor	6	7	8	9	10
Percent Variance	2.465	2.333	2.227	2.173	2.056

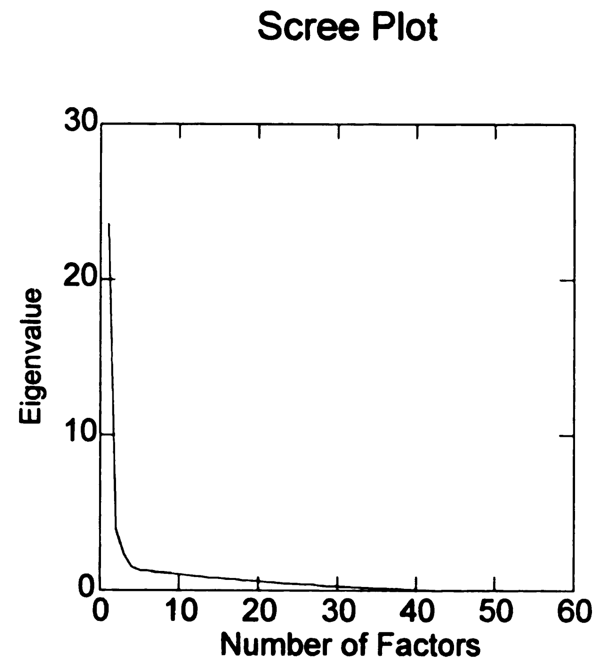


Figure 10: R91-1b Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	14.80	3.70	3.15	2.75	2.56
Factor	6	7	8	9	10
Eigenvalue	2.37	2.11	2.01	1.76	1.71
Factor	11	12	13	14	15
Eigenvalue	1.55	1.50	1.43	1.40	1.26
Factor	16	17	18		
Eigenvalue	1.11	1.05	1.03		

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	29.02	7.26	6.17	5.39	5.02
Factor	6	7	8	9	10
Percent Variance	4.65	4.14	3.94	3.45	3.35
Factor	11	12	13	14	15
Percent Variance	3.04	2.94	2.80	2.74	2.46
Factor	16	17	18		
Percent Variance	2.18	2.06	2.03		

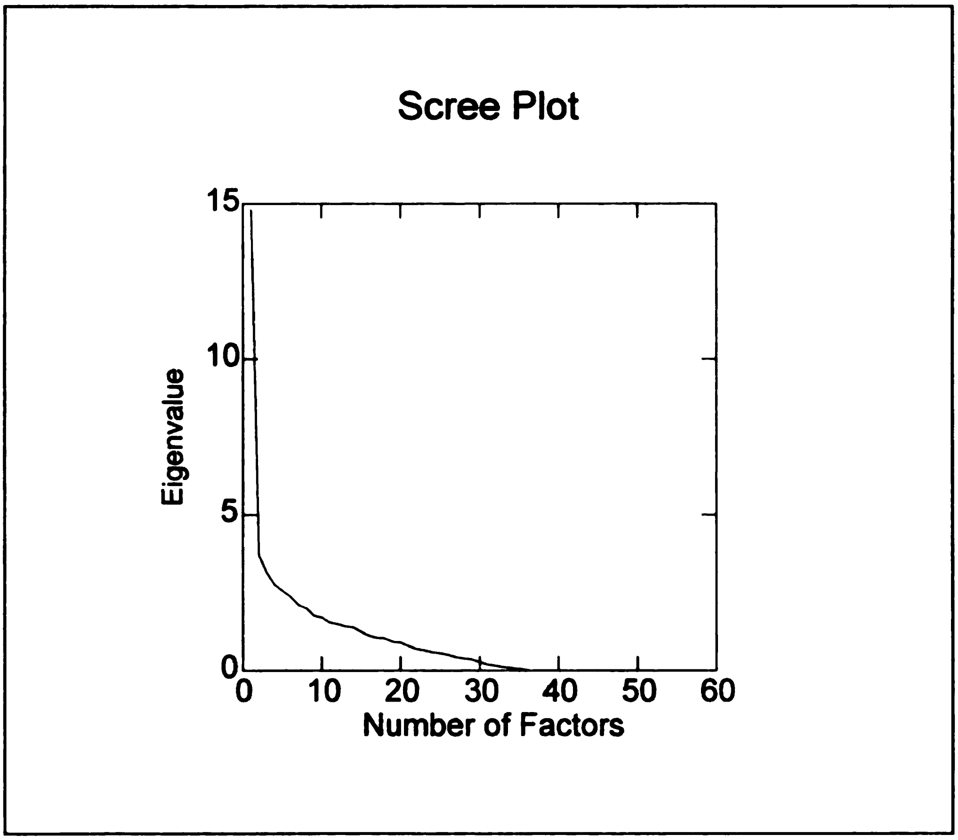


Figure 11: R91-2a Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	13.36	2.61	2.33	2.27	2.05
Factor	6	7	8	9	10
Eigenvalue	1.94	1.82	1.74	1.66	1.47
Factor	11	12	13	14	15
Eigenvalue	1.36	1.29	1.26	1.17	1.15
Factor	16				
Eigenvalue	1.10				

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	27.26	5.33	4.76	4.62	4.19
Factor	6	7	8	9	10
Percent Variance	3.96	3.71	3.56	3.39	2.99
Factor	11	12	13	14	15
Percent Variance	2.78	2.63	2.57	2.38	2.36
Factor	16				
Percent Variance	2.24				

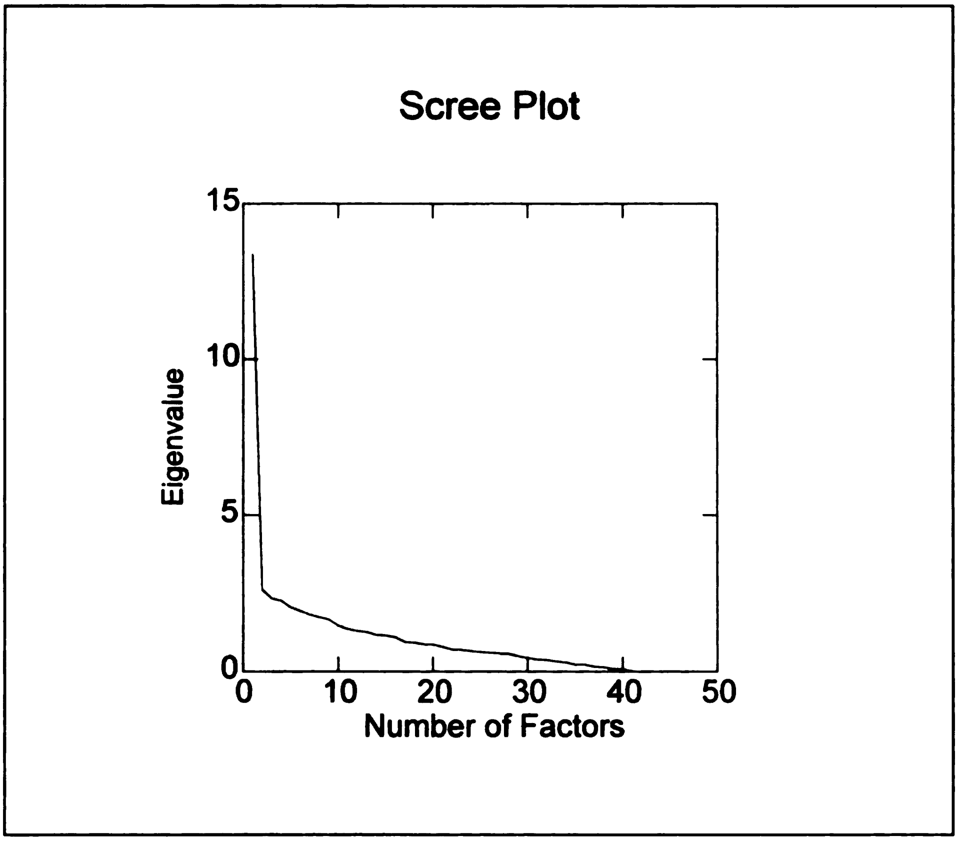


Figure 12: R91-2b Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	13.80	2.80	2.50	2.05	1.94
Factor	6	7	8	9	10
Eigenvalue	1.66	1.64	1.52	1.50	1.48
Factor	11	12	13	14	15
Eigenvalue	1.38	1.30	1.23	1.16	1.10
Factor	16				
Eigenvalue	1.06				

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	28.17	5.72	5.09	4.19	3.96
Factor	6	7	8	9	10
Percent Variance	3.39	3.35	3.10	3.07	3.02
Factor	11	12	13	14	15
Percent Variance	2.81	2.66	2.51	2.37	2.25
Factor	16				
Percent Variance	2.16				

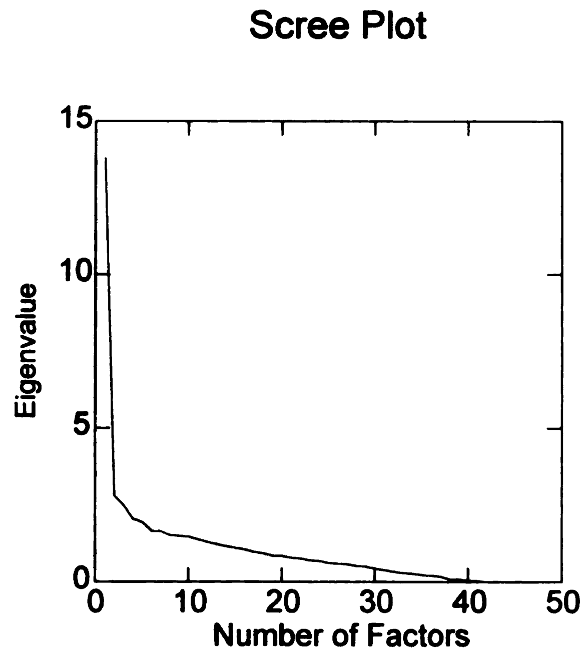


Figure 13: R91-3a Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	12.57	3.60	2.67	2.48	2.38
Factor	6	7	8	9	10
Eigenvalue	2.17	1.99	1.97	1.82	1.65
Factor	11	12	13	14	15
Eigenvalue	1.57	1.49	1.44	1.41	1.35
Factor	16	17	18	19	20
Eigenvalue	1.30	1.203	1.139	1.108	1.035

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	23.72	6.80	5.05	4.69	4.48
Factor	6	7	8	9	10
Percent Variance	4.10	3.75	3.72	3.43	3.10
Factor	11	12	13	14	15
Percent Variance	2.95	2.81	2.71	2.67	2.54
Factor	16	17	18	19	20
Percent Variance	2.45	2.27	2.15	2.09	1.95

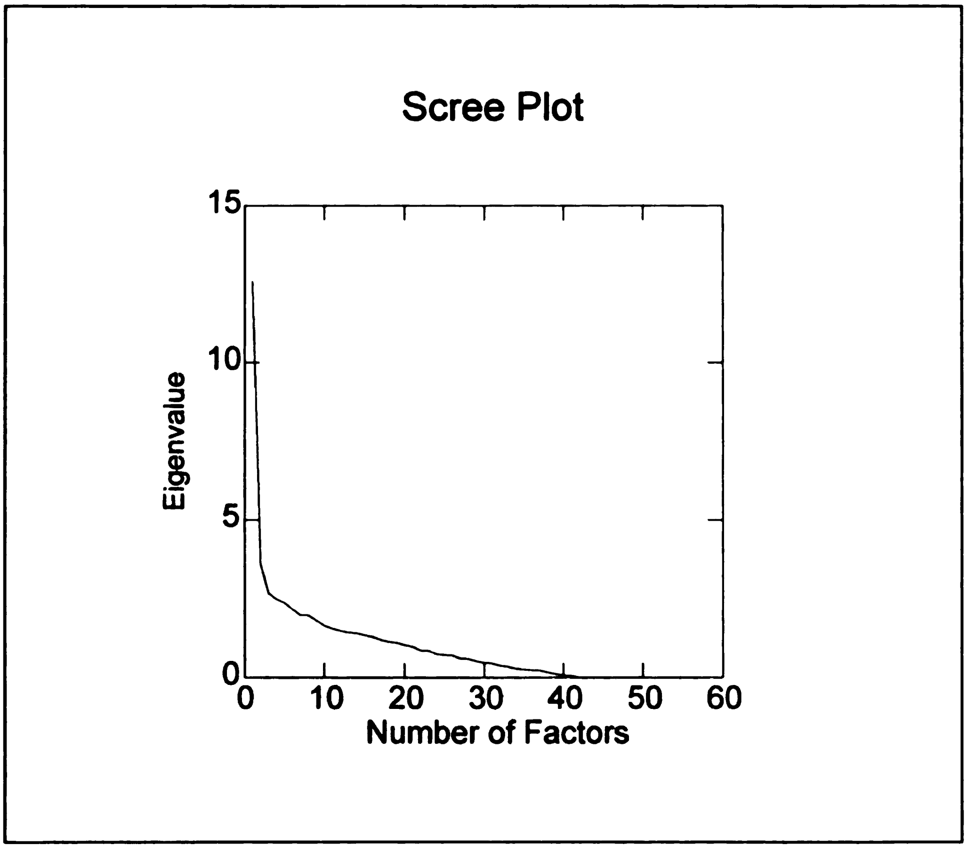


Figure 14: R91-3b Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	12.88	3.39	2.84	2.63	2.34
Factor	6	7	8	9	10
Eigenvalue	2.24	2.10	1.93	1.85	1.77
Factor	11	12	13	14	15
Eigenvalue	1.65	1.61	1.57	1.44	1.39
Factor	16	17	18	19	
Eigenvalue	1.35	1.21	1.138	1.072	

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	24.30	6.40	5.36	4.95	4.42
Factor	6	7	8	9	10
Percent Variance	4.23	3.96	3.64	3.49	3.34
Factor	11	12	13	14	15
Percent Variance	3.11	3.04	2.95	2.72	2.62
Factor	16	17	18	19	
Percent Variance	2.54	2.28	2.15	2.02	

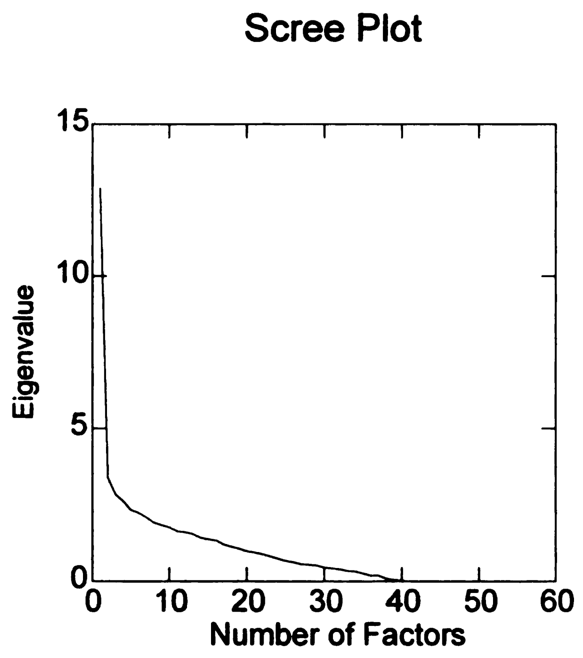


Figure 15: R92-1a Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	16.39	2.99	2.28	2.18	1.93
Factor	6	7	8	9	10
Eigenvalue	1.81	1.78	1.68	1.60	1.53
Factor	11	12	13	14	15
Eigenvalue	1.42	1.27	1.21	1.15	1.06
Factor	16				
Eigenvalue	1.01				

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	32.78	5.98	4.55	4.35	3.86
Factor	6	7	8	9	10
Percent Variance	3.63	3.56	3.36	3.21	3.07
Factor	11	12	13	14	15
Percent Variance	2.84	2.54	2.42	2.30	2.12
Factor	16				
Percent Variance	2.03				

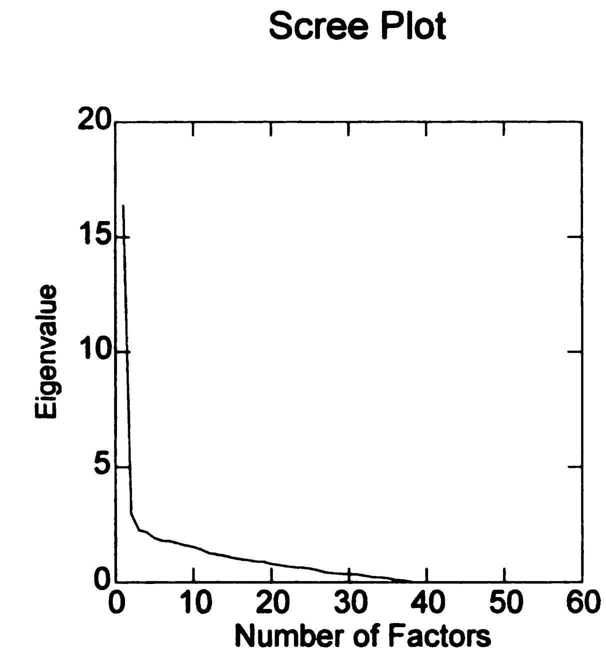


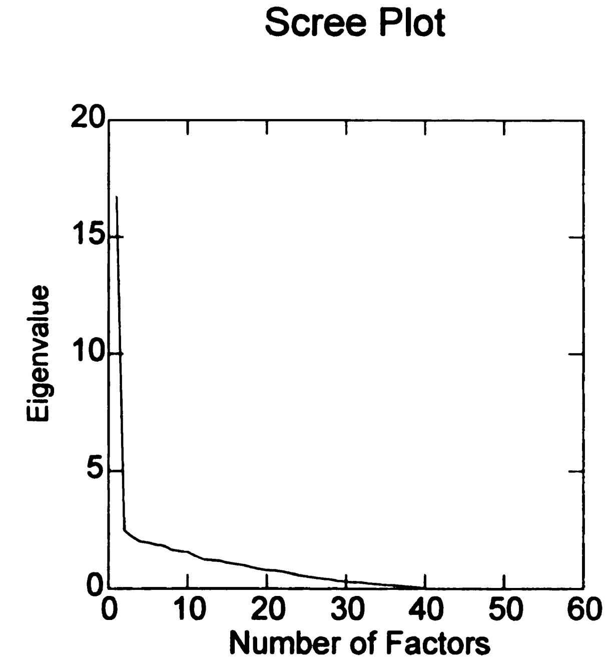
Figure 16: R92-1b Factor Analysis and Scree Plot Results

Variance Explained by Components with Eigenvalues Greater than One

Factor	1	2	3	4	5
Eigenvalue	16.74	2.48	2.22	2.02	1.96
Factor	6	7	8	9	10
Eigenvalue	1.87	1.83	1.66	1.61	1.58
Factor	11	12	13	14	15
Eigenvalue	1.40	1.26	1.23	1.20	1.11
Factor	16	17			
Eigenvalue	1.06	1.008			

Percent of Total Variance Explained

Factor	1	2	3	4	5
Percent Variance	33.48	4.97	4.43	4.03	3.93
Factor	6	7	8	9	10
Percent Variance	3.74	3.67	3.31	3.22	3.15
Factor	11	12	13	14	15
Percent Variance	2.80	2.52	2.46	2.40	2.22
Factor	16	17			
Percent Variance	2.11	2.02			



BIBLIOGRAPHY

Allen, M.J. and Yen, W.M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Anderson, N.J., Bachman, L., Perkins, K., and Cohen, A. (1991). An exploration study into the construct validity of a reading comprehension tests: triangulation of data sources. *Language Testing, 8*, 41-66.

Bachman, L.F. (1989). Response to Henning. *Language Testing, 6*, 223-229.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L.F., Davisons, F., Ryan, K., and Choi, I. (1994). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge: University of Cambridge Local Examination Syndicate.

Bachman, L.F., Kunnan, A., Vanniarajan, S., and Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability of two EFL proficiency tests. *Language Testing, 5*, 128-159.

Bachman, L.F. and Palmer, A.S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Barnwell, D.P. (1996). *A History of Foreign Language Testing in the United States: from its beginnings to the present*. Tempe, AZ: Bilingual Press/Editorial Bilingue.

Bensoussan, M. (1984). A comparison of cloze and multiple-choice reading comprehension tests of English as a foreign langauge. *Language Testing, 1*, 101-104.

Blais, J. and Laurier, M.D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing, 12*, 72-98.

Brown, J.D. (1995). *The Elements of Language Curriculum: A systematic approach to program development.* Boston, MA: Heinle & Heinle.

Brown, J.D. (1996). *Language Testing: A practical guide to proficiency, placement, diagnostic, and achievement testing.* New York: Regents/Prentice-Hall.

Bryk, A.S. and Raudenbush, S.W. (1988). *An introduction to HLM: Computer Program and User's Guide* (2^{nd} ed.). Chicago, IL: University of Chicago Department of Education.

Buck, G. (1990). The testing of second language listening comprehension. Unpublished dissertation, University of Lancaster, U.K.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8*, 67-91.

Carroll, J. B. (1983). Psychometric theory and language testing. In J.W. Oller (Ed.), *Issues in Language Testing Research.* Rowley, MA: Newbury House Publishers, 80-107.

Chappelle, C.A. and Abraham, R.G. (1990). Cloze method: what difference does it make? *Language Testing, 7*, 121-146.

Choi, I. and Bachman, L.F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing, 9*, 51-78.

Cook, H.G. (1995). A Hierarchical Linear Modeling approach to Assessing language growth. Unpublished Apprenticeship Paper.

Cook, H.G., Dunsmore, C.J. & Tan, H.S.S. *(1998)*. *Language Testing Video Series #1 Workbook.* Michigan State University Press: East Lansing, MI.

Chappelle, C.A. and Abraham, R.G. (1990). Cloze method: what difference does it make? *Language Testing, 7*, 121-146.

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Orlando, FL: Harcourt Brace Jovanovich.

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing, 1*, 50-69.

Davies, S. and West, R. (1989). *The Longman Guide to English Language Examinations.* Essex, UK: Longman Group UK Limited.

de Jong, J.H.A.L. (1984). Testing foreign language listening comprehension. *Language Testing, 1*, 97-100.

Dubin, F. and Olshtain, E. (1986). *Course Design: Developing programs and mateirals for language learning.* Cambridge: Cambridge University Press.

Duncel, P. (1993). Listening in the second/foreign language: Toward an integration of research and practice. In S. Silberstein (Ed.) *State of the Art TESOL Essays.* Alexandria, VA: TESOL, Inc.

Educational Testing Service. (1998). *TOEFL Test and Score Manual.* Princeton, NJ: Author.

Farhady, H. and Keramati, M.N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing, 13*, 191-207.

Frank, K. and Seltzer, M. (1990, April). Using the hierarchical linear model to model growth in reading achievement. A paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Freedle, R. and Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing, 10*, 133-170.

Freedle, R. and Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*, 2-32.

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

Fulcher, G. (1997). An English language placement test: issues in reliability and validity, *Language Testing, 14*, 113-139.

Hale, G.A. and Courtney, R. (1994). The effects of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing, 11*, 29-48.

Hale, G.A., Stansfield, C.W., Rock, D.A., Hicks, M.M., Butler, F.A., and Oller, J.W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6*, 47-76.

Harley, B., Cummins, J., Swain, M. and Allen, P. (1990). The nature of language proficiency. In B. Harley, J. Cummins, M. Swain, and P. Allen (Eds.), *The Development of Second Language Proficiency*. Cambridge: Cambridge University Press.

Hambleton, R.K. and Swaminathan, H. (1989). *Item Response Theory*. Dordrectht, The Netherlands: Kluwer Academic Publishers.

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, NJ: Sage Publications.

Henning, G. (1982). Growth-referenced evaluation of foreign language instructional programs. *TESOL Quarterly, 16*, 467-477.

Henning, G. (1989a). Meanings and implications of the principle of local independence. *Language Testing, 6*, 95-108.

Henning, G. (1989b). Comments on the comparability of TOEFL and Cambridge CPE.

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

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Huttenlocher, J.E., Haight, W., Bryk, A.S. and Seltzer, M. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236-249.

Jensen, C. and Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing,12,* 99-120.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14,* 23-46.

Larsen-Freeman, D. (1993). Second Language Acquisition Research: Staking out the territory. In S. Silberstein (Ed.), *State of the Art TESOL Essays: Celebrating 25 years of the discipline*. Alexandria, VA: Teaching English to Speakers of Other Languages, Inc. (TESOL).

Larsen-Freeman, D. and Long, M. (1991). *An Introduction to Second Language Acquisition Research.* New York: Longman, Inc.

Long, M. (1990). The least a second language acquisition theory needs to explain. *TESOL, Quarterly, 24:*4, 649-666.

Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, N.J: Earlbaum.

Lord, F.M. (1983). Small N justifies Rasch methods. In D.Weiss (Ed.), *New Horizons in Testing*. New York: Academic Press.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Language Testing, 10*, 211-234.

McNamara, T.F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*, 139-159.

Mehrens, W.A. and Lehman, I.J. (1991). *Measurement and Evaluation in Education and Psychology.* Fort Worth, TX: Holt, Rinehart and Winston, Inc.

Meisel, J.M., Clausen, H., and Pienemann, M. (1981). On determining developmental stages in second language acquisition. *Studies in Second Language Acquisition, 3*, 109-135.

Mellow, J.D. and Reeder, K. F. (1996). Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition, 18*, 325-350.

Millman, J. (1997). *Grading Teachers, Grading Schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press, Inc.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. J*ournal of Educational Measurement, 28*:2, 99-117.

Office of International Education Exchange. (Fall, 1993). *Annual Report of International Student and Scholar Enrollment: Michigan State University.* East Lansing, MI: Michigan State University.

Office of International Education Exchange. (Fall, 1997). *Annual Report of International Student and Scholar Enrollment: Michigan State University.* East Lansing, MI: Michigan State University.

Oller, J.W. (1979). *Language Tests at School: A pragmatic approach.* New York: Longman.

Oller, J.W. (1981). Language as intelligence? *Language Learning, 31*, 465-492.

Oller, J.W. (1984). Consensus and controversy. *Language Testing, 1*, 227-232.

Pearson, B.Z., Fernandez, S.C. and Oller, D.K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning, 43*, 93-120.

Pienemann, M., and Johnston, M. (1987). Factors affecting the development of language proficiency. In D. Nunan (Ed.) *Applying Second Language Acquisition Research.* Adelaide, Australia: National Curriculum Resource Centre.

Pienemann, M., Johnston, M. and Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10*, 217-243.

Perkins, K. and Miller, L.D. (1984). Comparative analyses of English as a Second Language reading comprehension data: Classical test theory and latent trait measurement. *Language Testing, 1*, 21-32.

Reckase, M.D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Rost, D.H. (1993). Assessing the different components of reading comprehension: Fact or fiction. *Language Testing, 10,* 79-92.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comperehension. *Language Testing, 1*, 147-170.

SPSS, Inc. (1996). *SPSS for Windows, Release 7.5.1.* Chicago, IL: Author.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrica, 55,*293-325.

Turner, C.E. (1989). The underlying factor structure of L2 cloze test performance in Francophone, university-level students: causal modelling as an approach to construct validation. *Language Testing, 6*, 172-198.

University of Michigan. (1996). MELAB Technical Manual. Ann Arbor, MI: English Language Institute, University of Michigan.

Wall, D., Clapham, C. and Alderson, C. (1994). Evaluating a placement Test. *Language Testing, 11*, 321-344.

Willms, J.D. and Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating schools effects and their stability. *Journal of Educational Measurement, 26*, 209-232.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design: Rasch Measurement*. Chicago: MESA Press.

Yalden, J. (1987). *Principles of Course Design for Language Teaching*. Cambridge: Cambridge University Press.