The Factor Structure of the English Language Development Assessment:

A Confirmatory Factor Analysis

by

Anju Kuriakose

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2011 by the
Graduate Supervisory Committee:

Jeff MacSwan, Co-Chair
Thomas Haladyna, Co-Chair
Marilyn Thompson

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

This study investigated the internal factor structure of the English language development Assessment (ELDA) using confirmatory factor analysis. ELDA is an English language proficiency test developed by a consortium of multiple states and is used to identify and reclassify English language learners in kindergarten to grade 12. Scores on item parcels based on the standards tested from the four domains of reading, writing, listening, and speaking were used for the analyses. Five different factor models were tested: a single factor model, a correlated two-factor model, a correlated four-factor model, a second-order factor model and a bifactor model. The results indicate that the four-factor model, second-order model, and bifactor model fit the data well. The four-factor model hypothesized constructs for reading, writing, listening and speaking. The second-order model hypothesized a second-order English language proficiency factor as well as the four lower-order factors of reading, writing, listening and speaking. The bifactor model hypothesized a general English language proficiency factor as well as the four domain specific factors of reading, writing, listening, and speaking. The Chi-square difference tests indicated that the bifactor model best explains the factor structure of the ELDA. The results from this study are consistent with the findings in the literature about the multifactorial nature of language but differ from the conclusion about the factor structures reported in previous studies. The overall proficiency levels on the ELDA gives more weight to the reading

and writing sections of the test than the speaking and listening sections. This study has implications on the rules used for determining proficiency levels and recommends the use of conjunctive scoring where all constructs are weighted equally contrary to current practice.

DEDICATION

This work is dedicated to my parents for their never ending faith in me and for their unconditional support and encouragement.

ACKNOWLEDGMENTS

I wish to thank Dr. Jeff MacSwan for his support the last eleven years. He encouraged me to apply for graduate studies in the field of language and literacy and was instrumental in getting funding support which meant a lot to me as an international student. He has always had my best interest in mind and has been consistent in encouraging me to finish this work that I started when I got mired down with family obligations. I am very grateful for having a caring, sincere, and passionate scholar as my mentor and chair. I wish to also acknowledge Dr. Haladyna who always was willing to spend time and resources with me and offered solutions to the problems I encountered in the last few years. I am very grateful for the insights he provided into the world of measurement and providing me with a data set that I could use for this analysis. Without his timely action and dedication to my project I would not have been able to finish this work. I also wish to thank Dr. Thompson who was generous with her time in answering my questions about structural equation modeling on a daily basis in my last semester. I thank her for advising me to do the analysis in the structural equation modeling framework and exposing me to this new methodology. Her support and willingness to help with the different issues that came up has helped tremendously to complete this work on time. I want to take this opportunity to thank Dr. Joe O'Reilly who was very supportive and gave me ample time off and

relieved me from my duties to complete this work. I am very grateful for the support and without having this time I would not have been able to graduate on time. Last but not least my husband, Shammi and our three kids (7 year old Nikhita, 4 year old Rahul and 1 year old Akhil) who put up with me when I worked for endless hours. I am so grateful for having a husband who travelled to India with the kids and stayed with them for the summer months so that I could finish this work.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

Introduction

Rationale

The purpose of this study is to examine the construct of English language proficiency (ELP) as measured by English language development assessment (ELDA), an ELP assessment used to measure language proficiency by several states in the United States (Arkansas, Iowa, Louisiana, Nebraska, South Carolina, Tennessee and West Virginia). ELP assessments are high stakes assessments for English language learners (ELL) because the scores are used for instruction, classification and promotion which affect their academic careers in school (Abedi, 2008). With the rising number of ELL students in our schools today, it is important that the ELL students are identified using a valid and reliable assessment which would give them access to appropriate programs and instruction in schools and would provide the teachers valuable information about growth over time.

Rising Number of ELL Students

Issues regarding the instruction and assessment of English language learners (ELL) have gained momentum after the No Child Left Behind Act (NCLB) passed in 2001. ELLs constitutes a significant proportion of the students in the schools, and there has been a significant increase in the last decade. According to the National Clearinghouse of Second Language Acquisition (2011), from 1999 to 2009 the number of

English language learners increased from 3.5 billion students to 5.3 billion which is a growth of 51%. The rising number of English language learners in United States (U.S.) schools brings unique challenges to teachers, schools and districts. The sole purpose of identifying students as ELLs is to cater to their academic needs so that they have the same academic opportunities as native speakers of English.

NCLB Mandates Testing

Standardized assessment results became an integral part of accountability after NCLB (2001) mandated that all states develop an assessment system aligned to the state standards and required that all students be tested, including ELLs, who by definition are not proficient in English. NCLB (2001) uses the term Limited English Proficient (LEP) and defines an ELL student as an individual who

a) is age 3 to 21 years;
b) is enrolled or preparing to enroll in elementary or secondary school;
c) was not born in the U.S. or whose native language is not English;
d) is a Native American, Alaskan Native, or a resident of outlying areas;
e) comes from an environment in which a language other than English has had a significant impact on an individual's ELP;

f) is migratory and comes from an environment where English is not the dominant language; and

g) has difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient level of achievement to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society.

The intent of Title III under NCLB (2001) is that all ELL students become proficient in English, and this will enable them to have the same opportunities to learn like native speakers of English. The section 3102 of the English Language Acquisition, Language Enhancement, and Academic Achievement Act, states the purpose is:

(1) to help ensure that children who are limited English proficient, including immigrant children and youth, attain English proficiency, develop high levels of academic attainment in English, and meet the same challenging state academic content and student academic achievement standards as all children are expected to meet;

(2) to assist all limited English proficient children, including immigrant children and youth, to achieve at high levels in the core academic subjects so that those children can meet the same challenging state academic content and student academic

achievement standards as all children are expected to meet, consistent with section 1111(b)(1).

It is very evident in the language of the law that ELL students should be held to same content standards as the other students. Teachers and schools should have the same expectations for this group of students as native speakers of English and the programs implemented should support the needs of ELL students.

Testing ELL students in English. The different parts of NCLB clearly indicate that ELL students should be tested every year on achievement tests as well as using the language proficiency test chosen by the state. Section 3102 (8) of the English Language Acquisition, Language Enhancement, and Academic Achievement Act, states that the purpose is:

> To hold state educational agencies, local educational agencies, and schools accountable for increases in English proficiency and core academic content knowledge of limited English proficient children by requiring
>
> 1. demonstrated improvements in the English proficiency of limited English proficient children each fiscal year; and
> 2. adequate yearly progress for limited English proficient children, including immigrant children and youth, as described in section 1111(b)(2)(B);

To satisfy the requirements outlined above, the test scores of all students including ELL students should be included for the accountability requirement for the Adequate Yearly Progress (AYP). Under NCLB, AYP is the evaluation that measures whether all students are making progress in attaining proficiency by 2014. Proficiency for AYP is defined as passing the achievement test aligned to the state standards. The percentage required to pass the test differs by grade and subject and the targets increase so that by 2014 the target would be 100% proficiency.

NCLB not only mandated achievement testing for all students, but also required that all ELL students be tested using a language proficiency test every year. One of the major requirements of NCLB is that the reclassification rate of ELL students should increase every year in all states. Reclassification rate is the rate of students identified as proficient in English based on the ELP test. States were required to set targets showing increases from year to year as part of federal accountability.

Testing language proficiency. As mentioned above, NCLB (2001) implemented new policies for ELLs. NCLB requires that all ELL students be tested using an ELP test every year, and that their progress in language development be monitored and reported as part of the accountability. Before this mandate was implemented, states were allowed to use different types of tests for testing English language proficiency and reclassifying ELL students as fluent English proficient. Some states used academic achievement tests for making this determination (Mahoney &

MacSwan, 2005). The states were not required to mandate a single test for comparability of scores from different school districts. Before 2005, states had different tests that test English Language Proficiency within the state. Most states had an approved list of tests that the district could choose for testing ELL students in the schools. The problem was that different tests were based on different theories and different standards, and hence the scores were not comparable. Therefore, a student could move from one district to the other and based on the ELP test that was chosen by the district, the student got different scores and different results. There was inconsistency in ELL classification across and within states (Abedi, 2004). Some states used ELP tests to make decisions, whereas some states used achievement tests to make these decisions (Linquanti, 2001).

The survey (Mahoney & MacSwan, 2005) done on the use and type of language proficiency assessments indicated that the most commonly used primary language assessments were the Language Assessment Scale–Español, (LAS[S]) in 11 states, the IDEA Language Proficiency Test–Spanish (IPT[S]) in 10 states, and the Woodcock-Muñoz Language Survey (Woodcock-Muñoz)[S]) in 5 states. These tests were developed by different testing companies independent of each other and had different levels of proficiency in the score reports. Based on the information from the technical manuals, the tests were not comparable

and the difficulty of the items of the test differed between the tests which made the scores incomparable.

States have implemented different policies regarding the instruction and assessment practices for ELL students (Mahoney & MacSwan, 2005). The method and the policies implemented in different states for the initial identification of ELL students is not a standardized process. Abedi (2008) states that the classification system that designates a student as ELL or English proficient is 'vague' and points out that it lacks strong theoretical foundations to make high stakes decisions. The use of ELP tests and the constructs measured by the tests and the validity of the scores have been of interest to researchers due to the differences in definition of ELP (Del Vecchio & Guerrero, 1995; Valdes & Figueroa, 1994; Zehler, Hopstock, Fleischmann, & Grenuik, 1994). The identification of the students as ELLs or non-ELLs is critical in determining the type of service that the students receive in school.

***Post NCLB language proficiency tests*.** NCLB (2001) under Title III mandated states to comply with the following requirements.

1. Each state had to adopt a single ELP test.
2. The ELP test had to be aligned to ELP standards adopted by the state.
3. The ELP standards had to be aligned to the content areas.

The mandate was that each state had to adopt a single ELP test that is aligned to the ELP standards to identify and reclassify ELL

students. Most of the states were not in a position to implement this change (Wolf et al., 2008). States were under pressure to get an off-the-shelf test to accommodate this requirement. In 2001, most of the states did not have ELP standards, and because of this, most states were not in a position to adopt an ELP test that was aligned with the ELP standards. In the survey done by Wolf et al. (2008), there were forty-three states in which the ELP test in use when the survey was done had not been used for more than five years. The mandate of using an ELP test aligned to ELP standards was a difficult task to accomplish in the short time span, and hence some states joined consortiums to develop the test and other states decided to buy one that was already available. The states that decided to use an off-the-shelf test had to augment it as soon as they had the standards developed to satisfy the requirement that the ELP test be aligned to the ELP standards. Forte (2007) reports that of the thirty-three states that responded to a survey about ELP tests, 26 states used off-the-shelf tests in 2005, but only seven states were using the same tests in 2007.

The requirement set forth by NCLB clearly stated that this test should test four different modalities in language – listening, speaking, reading and writing. This created a major shift in ELP testing where the types of items differed, and the definitions of language proficiency were different (Zehler et al., 1994). The requirements set forth by NCLB were prescriptive about the content that should be incorporated into the test.

They specified that the test of English language proficiency should include academic topic areas of mathematics, science, and social studies as well as topics that were related to the school environment (Fast, Ferrara & Conrad, 2004).

To summarize, the ELP tests that were mandated by NCLB had the following features. They were: (a) based on ELP standards; (b) also aligned to academic content standards; (c) considered secure and high stakes assessments; (d) focused on academic English; (e) inclusive of an oral language component which consisted of listening and speaking; (f) suited to provide a comparability of scores across grades to measure growth; and (g) tiered within grade levels (Abedi, 2007).

English language development assessment (ELDA). ELDA is an ELP test developed to satisfy the requirements of NCLB and it was developed in collaboration with multiple organizations. Based on the technical report (2005) the design, development and implementation of the ELDA was headed by the Council of Chief State School Officers (CCSSO) along with the participating states in the States Collaborative on Assessment and Student Standards for Limited English Proficient Students (LEP- SCASS). The members in CCSSO/LEP -SCASS include state education agency staff that combines their resources to develop assessment related projects that help member states. The LEP-SCASS consortium is composed of member states interested in developing resources for ELL students.

The CCSSO/LEP-SCASS solicited proposals for the development of The ELDA and American Institutes of Research (AIR) was selected to work collaboratively on the development of the ELP Assessment. AIR developed the items and the forms and the Center for the Study of Assessment Validity and Evaluation (C-SAVE) at the University of Maryland provided the research reports on reliability and validity (Lara et al., 2007).

The technical report (2005) produced by AIR clearly states that the driving force behind the construction of the ELDA was the six requirements specified in NCLB (2001): states must measure proficiency and show progress; assess all ELL students; independently measure the four skill domains of reading, writing, speaking, and listening; report a separate measure for comprehension; assess proficiency in academic language and in the language of social interaction; and align the assessments with their state Language Development (ELD) Standards. AIR constructed a set of core ELD standards based on the standards from the participating states and the LEP-SCASS approved them. These set of approved standards were used as the basis for test design and item development.

ELDA measures academic English as prescribed by NCLB. The construct was defined by the test developers as falling into two categories: the language used to convey curriculum based academic content and the

language of social environment of a school. Lara et al. (2007) describe the construct of academic English as measured by ELDA as,

> The concept of academic English is evolving, and it is important to make the point that although the ELDA items and prompts are written in the language of the classroom and of the academic subjects listed below, items do not require skills in or knowledge of content in those subjects. The concepts are not being assessed; the students' understanding of spoken and written texts about the concepts and their ability to write and speak about the concepts are being assessed. Any content a student is expected to use is provided in the stimuli or item prompt. (p. 48)

Validity of English Language Proficiency Score

The *Standards for Educational and Psychological Testing* (1999) states that "validity can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed uses" (p. 9). The first step involved in test development is to name the construct and define the construct. One of the biggest challenges in validating the English language proficiency score is that there is no consensus on the definition of this construct. Linguists have primarily defined language as a construct in terms of linguistic competence which refers to the use of language in a context. It does not involve literacy skills like reading and writing.

English language proficiency (ELP) tests have four discrete parts--reading, writing, speaking, and listening. This is a huge disconnect from the literature that attempts to define English language proficiency. The standardized language proficiency tests after NCLB got implemented in grades kindergarten to grade 12 test 'academic English'. However, the definitions were different and it lacked specificity and the standards that were used in the development of the test did not address the complex and multifaceted nature of language (Wolf et al., 2008). NCLB mandated that ELP tests measure the four discrete skills (reading, writing, listening and speaking) and referred to the construct measured as ELP. In other words, post-NCLB, the construct of ELP or academic English was operationalized by four different tests – speaking, listening, reading and writing.

The construct measured is of utmost importance in testing, and the definition of the construct drives the initial phases of test development. In practice, a construct is defined based on a theory, and a test is operationalized based on the definition of the construct. In this case, the ELP test is constructed based on standards, and the standards are not driven by second language acquisition theory. As mentioned above, ELP standards had to be aligned to the content standards, and with this mandate in place, ELP testing has shifted from the basic premise in measurement where the construct being measured is well defined and grounded in theory. Wolf et al. (2008) reports that "academic English constructs are described by listing tasks that occur in academic settings

without specifying academic language features." Academic English is not a construct that is well defined in the literature where the levels of proficiency can be clearly outlined. For example, Bailey and Butler (2003) defined academic language as "language that stands in contrast to the everyday informal speech that students use outside the classroom environment."

The focus of ELP testing as defined by NCLB is on measuring the academic English, but concerns remain among researchers whether ELP tests should be focused on language of academic content areas since there is no difference in the language used in different content areas (Abedi, 2008). The difference is in vocabulary, but not in the basic elements of language such as syntax, morphology, and phonology.

The construct of ELP as defined by NCLB mandates testing the four domains of reading, writing, listening and speaking. Each of these tests is administered separately and reports a separate score and a proficiency level. The test also reports an overall proficiency level which combines the four tests based on the rules adopted by the state. States were given the discretion to make the decision about how scores should be combined to create the overall proficiency level. States have used different rules in determining the overall proficiency levels. The rules used by states are different for the determination of overall proficiency levels (Porter & Vega, 2007). The two models used for scoring are conjunctive and compensatory. Conjunctive scoring means that the student has to be

proficient in all four domains to be considered proficient whereas in the compensatory method, doing well on one domain can compensate for the lack of performance in the other domains. For example, the World-Class Instructional Design and Assessment (WIDA) consortium developed an ELP test (ACCESS for ELLs) based on the NCLB mandates and the scores are weighted (15% speaking, 35% reading, and 35% writing) to determine the overall score. Reading and writing comprise 70% of the overall score, and accordingly, may compensate for the lack of performance in listening and speaking (Bauman, Boals, Cranley, Gottlieb, & Kenyon, 2007). ELDA uses a different weighting rule where the final proficiency level is determined by combining the proficiency levels from each of the domains. The reading and writing proficiency levels on the ELDA contribute more to the overall proficiency level than speaking and listening (Lara et al., 2007). A detailed explanation of how ELDA determines proficiency levels is provided in the section that provides the description of the instrument.

This study examined the factor structure of ELP as measured by the English Language Development Assessment (ELDA), an ELP test that was aligned to the ELP standards based on the NCLB mandates described above. This study addressed whether ELP is a unidimensional or multidimensional construct. The use of a single combined score from the four domains of listening, speaking, reading, and writing suggest that this construct may be regarded as unidimensional. However, the four

different skills are tested separately and can be considered as four independent constructs related to each other. This study aims to provide empirical validity evidence that will allow test users more insight into the factor structure of ELP as measured by ELDA. The study also addressed whether the factor structure is the same for students in grades three to five and in grades 9 to12. The research questions addressed in this study were:

1. Which model best represents the factor structure of the ELDA with the four language arts abilities (reading, writing, listening, and speaking)?
    i. Is the ELDA represented well by a factor structure that includes the four hypothesized factors?
    ii. How does the fit of the hypothesized four-factor model compare to a one-factor model of English language proficiency?
    iii. Is the second-order model a good fit for the data?
    iv. Can the bifactor model explain the structure of the ELDA?
2. Is there a difference between the factor structure for students in grades three to five and students in grades 9 to 12?

The next section provides a review of the literature. The topics covered include first and second language acquisition theories, a brief section about the impact of NCLB in the development of ELP tests,

validation studies conducted on ELP tests after NCLB was implemented, and construct validation studies done on ELP tests are also reviewed.

Chapter 2

Literature Review

This chapter covers a review of literature on different topics that are important to this study. The first part of the review is about first and second language acquisition theories. The theories of language acquisition are discussed because this study is about testing ELLs on the construct of ELP. The second part of this chapter is focused on testing ELP and the challenges involved with that in the light of regulations implemented by NCLB. The third section is a brief overview of validation because this study provides empirical validity evidence on the construct of ELP. This section also talks about factor analysis and why this method is effective to answer the research questions. The last part of this chapter reviews validity studies and factor analytic studies of language proficiency tests and briefly addresses the findings and conclusions in the literature about ELP tests.

Language – The Innate Human Ability

Language is one of the distinct abilities of human beings and all human beings acquire a language irrespective of culture and socioeconomic status (Slobin & Bowerman, 1985; Pinker, 1994). This is a complex ability where there are multiple processes working together. Acquiring a language begins before a baby is born. Even though most people cannot articulate how they acquired their language, they all

inevitably master the language of the community in which they grew up. Bialystok and Hakuta (1994) summarize the complexity of language learning in this way:

> Even a brief moment of reflection reveals that language learning takes place in a complex ecology and not in a laboratory. The full repertoire of our human nature, ranging from our cognitive machinery to our social and communicative needs, is engaged in the activity. It will be overwhelmingly difficult and ultimately unproductive even to attempt to study a system of this complexity in its entirety. (p. 8)

The complexity of understanding how we acquire a language has made it difficult to define the construct and to measure it. There are numerous theories about language development and about how children acquire their first language. While much has been learned there remains little consensus about how children acquire language, and even less agreement about the definition of the construct of language proficiency in measurement contexts, perhaps due to the complexity of understanding the different processes involved in acquiring language.

All theories agree that children must be exposed to a language in order to acquire it. Originally it was thought that language was learned through imitation. Skinner's (1957) theory suggested that language learning followed from stimulus and response mechanisms which he had developed as part of his broader psychological theory, behaviorism.

Behaviorism emphasized that language learning occurred through imitation and repetition. Children imitate what they hear, and they become proficient speakers of language. This theory was limited and could not explain the acquisition of language because children acquire the rules of sentence structure without direct instruction and they have the ability to create sentences that they have never heard before (Chomsky, 1959; Hauser, Chomsky & Fitch, 2002). According to the behaviorist view, language was learned just like a person learns any other behavior through imitation and reinforcement, and hence it is no different than learning any other skill. This view was prevalent until Chomsky's (1959) critical view of Skinner's work which also projected an alternative view.

Chomsky (1959) revolutionized both linguistic theory and the theory of language acquisition by arguing that language was too complex and structural diverse be explained by a simple stimulus and response approach. Chomsky noted that children say things they have never heard before, what he termed the creative aspect of language, ruling out the notion that language acquisition was directly related to imitation. Chomsky's view of language is that it follows from an innate, species-specific ability found in all human beings. As Chomsky (1975) noted:

> A human language is a system of remarkable complexity. To come to know a human language would be an extraordinary intellectual achievement for a creature not specifically designed to accomplish this task. A normal child acquires this knowledge on relatively slight

> exposure and without specific training. He can then quite effortlessly make use of an intricate structure of specific rules and guiding principles to convey his thoughts and feelings to others, arousing in them novel ideas and subtle perceptions and judgments. (p. 4)

Chomsky's theory is referred to as generative grammar, and the premise is that there are a finite set of rules in any language that govern how sentences are made. In his view, children are born with an innate capacity to acquire language which he called the language acquisition device. All languages have rules and children are prewired with the innate implicit knowledge of language which in later years, came to be known as Universal Grammar (UG) (Chomsky, 1986, 1995). This view of language as innate is supported by the evidence that all children acquire very complex grammatical structures at a very early age, and they produce sentences to which they have never been exposed in their environment. The number of responses a child can construct is infinite, and they create complex grammatical structures in their first language. This ability in human beings has led scholars in the field to believe that language acquisition is an ability governed by innate principles of UG. All typically developed human beings have this innate ability to acquire the language, and they specifically acquire the language of their speech community. There is variation in language use across different speech communities, but central to linguistics today is the premise that all languages are equally

rich and complex (Crystal, 1986; MacSwan, 2000; MacSwan & Rolstad, 2010).

Second Language Acquisition Theories

While work in mainstream linguistics has focused on language structure and first language acquisition, a number of researchers have also addressed the important question of how we acquire a second language. There are many theories that try to explain the nature of second language acquisition. One of the main distinctions seen in the literature is the difference between acquiring and learning a language, as originally stressed by Krashen (1982). For Krashen acquisition refers to the natural process in which children acquire their first language. They acquire the language when they are exposed to it in real life situations. When children acquire a language they are not conscious about the overt grammatical rules. However, Krashen argued that *learning* a language is different from *acquiring* a language. Learning a language is a very conscious process in which you learn the rules, parts of speech, subject verb agreement, and even the rules of language use in social contexts. Krashen's distinction between learning and acquiring was used to argue that second language acquisition could, in principle, model first language acquisition, rather than require language learning, if only the learning environment and context are appropriately constructed. While controversial, the theory and basic distinctions remain highly influential in the field of second language acquisition. In addition, researchers have been concerned with the effects

of language acquisition on school subject matter learning, and have therefore raised questions about how language acquisition and bilingualism might be related to cognitive abilities.

For instance, there have been studies that suggest that bilingualism, or being proficient in two languages, has cognitive advantages (Duncan & DeAvila, 1979; Kessler & Quinn, 1982; Bialystok & Martin, 2004). However, other studies suggest that there are negative consequences based on the relative levels of proficiency developed in the first and second languages (Cummins, 1979). Cummins (1979) hypothesized the notion of 'semilingualism' claiming "there is strong evidence that some groups of minority language and migrant children are characterized by 'semilingualism,' i.e. less than native like skills in both languages with its detrimental academic and cognitive consequences" (p. 228). Cummins is frequently referenced in the literature in the literature for his threshold hypothesis which is about the relationship between cognition and bilingualism, and the phrases he coined to describe compartmentalized language skills, basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP).

The threshold hypothesis. The premise of the threshold hypothesis (Cummins, 1979) is that there are three different levels of language competence in the first and second language. Based on the level of language competence in the first and second language, there can be positive or negative effects. The bottom level is described as the level

in which the child has low levels of mastery in the first and second language. They are limited in their ability to use both languages and this has negative cognitive effects. In the second level, the child is described as being proficient in one, but has limited ability in the second language. This level has no negative or positive consequences. In the third level which is the highest level, the child is proficient in both languages with balanced ability in this case, hence called "balanced bilinguals." At this level, children have positive cognitive advantages compared to monolinguals. The threshold hypothesis also presupposes that if children are exposed to second language without being proficient in the first language, they become "semilingual," which means they are not proficient in either language. If a person is proficient in a language, there are no negative cognitive effects. If a person is proficient in both languages, then there are positive cognitive effects.

This theory has been criticized in the literature. The threshold hypothesis, as proposed, is a deficit theory and semilingualism has been referred to as a half-baked theory (Martin-Jones & Romaine, 1986). MacSwan (2000) refutes the basic claims underlying the semilingualism thesis by addressing the types of evidence that were advanced in support of this idea and concludes that there is no empirical evidence that such a state exists. There is also no evidence in the literature that a typically developing child exposed to language will not become proficient in a first

language. All native speakers of a language are proficient in that language, by definition.

MacSwan and Rolstad (2010) also note that semilingualism is not supported by relevant evidence to acknowledge and compared the idea to prescriptivism. Prescriptivism is the view that some languages are inherently better than others. MacSwan and Rolstad (2010) state that the idea of semilingualism is no different than prescriptivism because purported evidence of semilingualism relies on prescriptivist notions of language.

BICS and CALP. After the threshold theory, Cummins (1979) developed the "developmental interdependent hypothesis," the view that the development of the second language is dependent on the competence achieved in the first language. In order to explain language development, Cummins (1984) formulated another theory which distinguishes between surface fluency or basic interpersonal communication skills (BICS) and the more evolved language skills, cognitive academic language proficiency (CALP) that students need to perform well in school. BICS, as the name suggests, is about conversational skills or 'playground language' which children acquire early on when they are exposed to a second language. According to this theory, the interpersonal conversational skills that are acquired are less cognitively demanding. Cummins further claims that BICS provides the user context and the clues and support from facial expressions and body language provide support to understand the

language better. CALP is considered as the cognitively challenging and is associated with literacy. This academic language proficiency is considered as the superior skill in this theory. This is context reduced and requires second-order thinking skills. There has been a lot of criticism of this model (Edelsky et al., 1983; MacSwan, 2000; Romaine, 1995; Wiley, 1996) because it is a compartmentalized view of language, and most of the research on language acquisition and learning suggests that language learning is a very complex process that involves different cognitive and academic abilities that interact with each other. This view of language proficiency confounds language ability and academic achievement, and it does not take into account crucial differences between first and second language acquisition. This distinction suggests that schooling improves our language, which implies that the language of the educated is better than the language of the unschooled (MacSwan & Rolstad, 2003).

The distinction between BICS and CALP remains controversial and it is characterized in the literature as an oversimplification of complex language and cognitive processes. This distinction has face validity, but there is no empirical evidence that this distinction exists (Martin-Jones & Romaine, 1986). The definition of this partition in the development of the first language is not precise to be tested and hence cannot be validated. One of the major criticisms is that this distinction is value laden (Wiley, 1996) because in this frame work, BICS is seen as inferior and less cognitively demanding than CALP. The other distinction this theory makes

is that BICS happens first and is developed in a short period of time while CALP happens much later and as a result of schooling. The idea that CALP is much richer and complex has not been empirically validated, and there is no evidence that supports this claim. Even though this could be the order in which a child may learn the second language, this does not have to be in this order. Students learning a second language in college learn to read and write first, and it is much later that they learn to speak, if they ever become proficient in the second language. It also needs to be noted here that there is a lot of evidence that suggests speaking a language or communicative competence is cognitively demanding. Learning to communicate in a second language and using language effectively in social situations is a skill that a second language learner takes time to master, and hence this theory lacks evidence to support the claims made.

This is a deficit view, and there are negative consequences for ELL students because the BICS/CALP distinction does not clearly differentiate language proficiency from academic achievement. In this framework, CALP is developed later in life as a result of schooling. As Cummins (2000) puts it:

> In monolingual contexts, the [BICS/CALP] distinction reflects the difference between the language proficiency acquired through interpersonal interaction by virtually all 6-year-old children and the

> proficiency developed through schooling and literacy which continues to expand throughout our lifetimes. (p. 63)

In this framework, Cummins (2000) clearly states that schooling improves language and that CALP has “complex grammatical structures, greater demands on memory, analysis and other cognitive processes.” This assumes that the language of the educated people is superior and better than the unschooled. This hierarchy inherent in the BICS/CALP framework has made scholars describe it as a half-baked theory, compare it to prescriptivism and as a deficit theory (Edelsky et al., 1983; MacSwan, 2000; MacSwan & Rolstad, 2003; Martin-Jones & Romaine, 1986; Wiley 1996). Both the threshold hypothesis and the BICS/CALP distinction do not have any evidence that supports these claims. Both the theories are compared to prescriptivism and discussed in the literature as a deficit theory and explaining the language ability as semilingualism may have negative consequences if educators use these terms to describe and refer to second language learners in schools.

English Language Proficiency Testing – Policy and Relevance

The ELP tests as mandated in Title III (2004) are designed to measure ELP standards. ELP standards have to align to the academic standards in content areas. So in this process the theoretical construct of language proficiency was redefined by NCLB as language that is required to be successful in schools, or in other words “academic language.” The guidance given to each state as outlined in Title III was to:

> Describe how the agency will establish standards and objectives for raising the level of English proficiency that are derived from the four recognized domains of speaking, listening, reading, and writing, and that are aligned with achievement of the challenging state academic content and student academic achievement standards. (2004)

All states were required to develop ELP tests that were compliant with the mandate mentioned above. The construct being tested is referred to as language proficiency even though the requirements described the construct as the language that is used in the content areas or "academic language" (Cummins, 1981). The BICS and CALP theory as described by Cummins (1981) has been a controversial idea and NCLB has made this concept in to a law. The mandate states that not only conversational skills (listening and speaking) should be tested but academic English (reading and writing) has to be tested and should be used for identifying and reclassifying English language learners.

Most of the research in linguistics about language proficiency and language acquisition revolves around language ability as a communicative ability where the focus is on effective communication. Language and literacy are two different constructs. Based on the theories of language acquisition, acquiring a language is an innate ability whereas literacy is a very different skill. Literacy comes from direct instruction and a skill that is learned and in most children develops these skills as a result of schooling.

The construct of reading and writing is defined in the literature separately but there are no theories about how the discrete skills of listening, speaking, reading and writing combine together to form one construct called language proficiency. The ELP tests focuses on the concept of "academic English" and uses these language and literacy skills to describe language proficiency. There is a disconnect between theories in second language acquisition and second language testing in the K-12 setting. Testing language proficiency as described in the federal mandates combines the four domains of speaking, reading, listening and writing to create an overall proficiency level. States like Arizona combine the scores from these four subtests to create one score, a language proficiency score, to make decisions about entrance and exit from ELL programs. This is high stakes decision, especially in states with restrictive language policies like Arizona, California, and Massachusetts, where the mandate is English only (Gandara & Hopkins, 2010). The decision that the school/district makes about students based on the score determines whether the child will have access to core curriculum. The ELP score should be valid to make this decision and validation evidence has to be collected and analyzed to ensure the validity of the ELP score.

Validation

There are different forms of validity evidence that can be collected, but validity itself is unitary concept. There are different kinds of validity evidence that can be gathered to make the validity argument based on the

interpretation of the score. If the score is used for multiple purposes, then validity evidence has to be gathered for each purpose. So validity is not a function of the test itself but it integrates various kinds of evidence that will support the intended use of the score. It is also important to note that different test score use warrants collecting and analyzing different kinds of validity evidence to ensure that the use of the score is valid.

Different kinds of evidence can be collected to make the validity argument. Based on the *Standards for Educational and Psychological Testing* (1999), there are different kinds of evidence that should be assembled for a test is valid. Evidence that can be collected includes information that supports the validity, and evidence should be collected that threatens the validity of the test. The evidence that can be collected to support the validity argument include evidence based on test content, response processes, internal structure, item quality, score comparability, standard setting, other measures, consequences of testing, and supporting documentation. Evidence can also be collected that threatens validity. They include construct irrelevant variance and construct underrepresentation.

This study focuses on the internal structure of the test and how this adds to the empirical evidence. The evidence based on internal structure explores how the relationship among the items relates to the construct being tested. The construct being tested can be unidimensional or multidimensional. One of the most important steps when developing a test

is to name and define the construct that is being measured. The test developer should have a clear definition about the dimensionality of the construct being measured.

Dimensionality. Dimensionality is a term used in measurement to describe the number of abilities or constructs tested on a particular test. A test can have one dimension or multiple dimensions. Defining dimensionality of the test is one of the most important steps in test construction. A test can be unidimensional or multidimensional based on test content. The number of dimensions varies depending on the construct being measured. Tate (2002) defined test dimensionality as the minimum number of examinee abilities measured by the test items. Dimensionality refers to the abilities of the test taker. If there is only one score given for a test, the assumption is that the test is unidimensional which suggests that there is only one ability or construct that is being measured by that instrument. Unidimensionality in the strict sense assumes that the items on a test are strictly homogenous or, in other words, all items on the test measure the same construct, one common attribute or latent ability.

A single score implies that the test is unidimensional. In the development of a test, the test publisher has to define the construct being measured. In the definition there should be a clear indication as to whether the construct is unidimensional or multidimensional. Depending on the dimensionality of the construct, a decision has to be made about

subscores. Subscores come from different scores that can be added up in a meaningful way to measure a construct. For example, speaking and listening skill scores can be combined to have a measure of language competence if there is a theory that supports that construct. A score put together from different subscores is called composite score or total score.

More than one score or subscores on a particular test implies that there are different abilities being tested and each one of those parts merits a separate score. Such a test is multidimensional because there are multiple abilities being tested at the same time. For example, a mathematics word problem is multidimensional because it tests two types of abilities of the test taker. The test taker has to read and understand the problem and then have to perform the mathematical task described in the problem. If it is a single composite score that this produces, then the test is unidimensional. But if the construct being measured is mathematics ability, then the score is not a valid measure of mathematics ability. This is talked about in the literature as causing construct irrelevant variance. The linguistic complexity of the mathematics item produces variance that is irrelevant to the ability being measured.

Tate (2002) states that the test dimensionality is the minimum number of dimensions or abilities required to explain all test related differences among the test takers. Ackerman (1989) describes this interaction between examinees and the items should be empirically tested, and that unidimensionality should never be assumed. When the abilities

of the test takers are different in the skills tested, and the items differentiate between levels of the skills, then the test should be considered multidimensional.

Empirical analysis should be done to confirm whether a test is unidimensional and that the single score represents a single statistical ability. This will provide conclusive evidence for the reliability estimates to be more accurate and that there is no construct irrelevant variance and bias in the items. This will also make the equating across years less complicated for score comparability. However, if the empirical analysis shows that there is more than one-factor, then, steps need to be taken to address the problem because the unidimensionality assumption is violated.

Validation using a composite score. The use of a composite score implies that multiple abilities are being measured. Subscores are derived from parts of a test and used to create the composite score. A composite score created with different subscores can be misused and misleading if the test user is not aware of how the composite is created. A test taker can perform well on a subtest and score really high and on another sub test can score really low. A composite score will not reveal these examinee abilities, and this can be misleading. For example, if a composite score is created from a reading score and a writing score, then the test user should be aware what the score means. When a composite score is used, a high score on one part can compensate for a lower score

on the other part. So in this example, the reading score can be high and writing score could be low, but a composite will not give adequate information about the individual subtest. During test construction, this has to be defined based on the construct that is being tested. Construct definition should include whether a single score is sufficient or the construct warrants different scores for each of the abilities measured. This is a crucial step in the development of any test because this determines how the test score is going to be interpreted and used which in turn adds to the validity of the test.

Dimensionality and validity. Dimensionality of the test has implications on validity. Tate (2002) describes test dimensionality as an integral part of addressing issues of test validity, reliability, fairness and score comparability. Haladyna and Downing (2004) discuss in detail why dimensionality should be carefully considered in test development. The important issues include reliability, the use of subscores, ensuring fairness or test bias, score comparability and standard setting to establish cut scores. The discussion below is the major highlights.

Reliability. There are other implications for validity that arise from the study of dimensionality. Internal consistency reliability will be underestimated if the item responses are multidimensional. Internal consistency reliability estimates will be lower if the item responses reflect multidimensionality. Internal consistency reliability will be underestimated if the item responses suggest a multidimensional structure.

The dimensionality of the test has to match the test content structure in order for a test to be valid. Tate (2002) suggests an empirical analysis of test structure should be conducted after the test plan is matched with test specifications or the test blue print. One of the major steps in the development of a standardized test is creating the test specifications. Test specification provides detailed information about the proportion of items that will be on the test based on each content domain.

The standard practices of computing reliability assume that the test is unidimensional. The literature on this topic suggests that this assumption is always violated to some extent. The test developer should check whether the estimate of reliability is accurate and whether the unidimensionality assumption is appropriate for the test content. In classical test theory, there is the assumption that items are locally independent and that the items are homogenous.

Use of subscores. The use of a total score or subscore may be affected by dimensionality. If subscores are going to be used, item analysis should be done on subscores because this will produce different results. The decision about the use of sub scores or total scores should be based on the construct being measured. If the subscores are highly correlated, then this may be considered as evidence that the test is unidimensional, and in that case, the use of subscores is not warranted. If the validity evidence suggests a multidimensional interpretation, then the subscores can be informative if they are reliable.

Score comparability. Score comparability over time is very important in large scale assessments. One of the major goals of large scale assessments is to compare scores from year to year to analyze long term trends. In order to compare scores from year to year, the tests must be equated, and this has to be done with careful consideration given to dimensionality of the test. The test structure must be maintained to compare the results over time or it may threaten the validity of the interpretation of the score. Score comparability allows test users to analyze growth and long term trends in test scores. This is very important because the test scores from longitudinal analysis are used for high stakes decisions which include teacher and principal incentives and school closures. Year to year comparisons will be meaningless if dimensionality is not considered while equating and scaling.

Setting cut scores. The way cut scores for different levels are set can be affected by dimensionality. If the test is designed such that subscores are used to make decisions, then each subtest will have cut scores. For example, for a language proficiency test, the speaking test will have a cut score to identify the different levels of speaking proficiency, and the listening subtest will have different levels of listening proficiency. If the scores from the listening part and the speaking part are combined to create a composite score for language proficiency, then a cut score has to set separately for this purpose. If the test uses a composite score, then there is the possibility that one or more sections of the test are given more

weight. This implies that a low score on one of more of the sections is not of concern and that the interpretation of the score remains the same irrespective of differences in the performance in the different parts of the test.

Methods of Assessing Dimensionality

Factor analysis is one of the common methods used to analyze the number of factors in a set of observed variables. This analysis assumes that the observed variables are linear combinations of factors. Factors can be defined as hypothesized, unmeasured, and underlying variables which are presumed to be sources of the observed variables (Kim & Mueller, 1978). Factor analysis is a set of correlational analysis designed to examine the relationship among different observed variables. This gives the smallest number of latent unobserved factors that explains the observed variables. Many definitions are offered in the literature for factor analysis. Reymont and Joreskog (1993) define factor analysis as a:

> Generic term used to describe a number of methods designed to analyze interrelationships within a set of variables or objects that account for the construction of a few hypothetical variables called factors, that are supposed to contain the essential information in a larger set of observed variables which reduces the overall complexity of the data by taking advantage of inherent interdependencies. (p. 71)

Factor analysis was first developed by Spearman in 1904. Factor analysis helps the researcher the correlation and covariance in a set of observed variables by a set of unobserved latent factors. The factors can be common or unique. A common factor affects more than one observed variable and a unique factor affects only one of the observed variables. There are two major kinds of factor analysis – exploratory and confirmatory. The decision about the one to choose is based on the purpose of the analysis.

Exploratory factor analysis. Exploratory factor analysis (EFA) is generally used when the researcher does not have a priori hypothesis about the number of latent factors. This analysis allows the researcher to explore the underlying factor structure. This helps the researcher to check whether the construct intended is measured by checking whether the scores on the test accurately measures what it is supposed to be measuring. The purpose of EFA is in this sense to come up with a theory about the factor structure of the underlying data.

Interpreting the results from an EFA is difficult because the researcher does not have enough knowledge about the factor structure. The common model used is a linear model and it may not fit all data. Causal relationships tend to be nonlinear and forcing a linear relationship may yield misleading results. The results sometimes may not be meaningful because the method gives the researcher the best fit. The

results are also hard to interpret because factor structure results are driven by the method and the rotation procedures used.

Mulaik (1972) stated that:

> In a practical sense, there is no question that EFA serves a useful purpose in suggesting hypotheses for further research. But one must not be misled into thinking that EFA- or any exploratory statistical technique, for that matter-is the only way, or even the optimal way, available to us to obtain suggestions for hypotheses. One's own direct experience with a phenomenon often suffices to suggest hypotheses. (p. 269)

According to Fabrigar, MacCallum, Strahan, and Wegener (1999) the researcher has to make five methodological decisions when conducting a factor analysis. They are (a) the measures to include in the study, (b) to determine whether EFA is the most appropriate method to answer the research question, (c) a factor extraction procedure should be chosen to answer the research question (d) a decision has to be made about the number of factors that should be extracted and (e) a decision has to be made about the rotation that will be used (orthogonal or oblique).

The main purpose of EFA is to find the minimum number of factors, or latent constructs, that can account for the relationship among the measures but the researcher has to be knowledgeable about each of the decisions that have to be made, and if enough attention is not given to the

details about the choices, the results may be inaccurate (Fabrigar et al., 1999).

Confirmatory factor analysis. Confirmatory factor analysis (CFA) is a type of structural equation modeling where the researcher hypothesizes the number of factors and the relationship of factors to all the measures. In CFA the researcher starts with a theory about which factors are correlated with the different variables or items based on the construct being tested. In this analysis the researcher has the ability to explicitly check the factor structure, because in this model, the number and composition of factors is predetermined. The analysis helps the researcher to check how well the factor structure explains the fit of the model.

The literature suggests that CFA is more appealing than EFA because the researcher is testing a priori hypotheses. The hypothesis is based on strong empirical evidence, and testing this enables the researcher to confirm the latent unobserved factors. CFA also allow the researcher to answer a wider range of research questions compared to EFA.

Commenting on the utility of CFA, Gorsuch (1983) noted that “CFA is powerful because it provides explicit hypothesis testing for factor analytic problems....CFA is the more theoretically important-and should be the much more widely used-of the two major factor analytic approaches" (p. 134). He specified that exploratory methods should be "reserved only

for those areas that are truly exploratory, that is, areas where no prior analyses have been conducted” (p. 134).

CFA allows the researcher to test multiple hypothesized models at the same time, and this is a big advantage. The researcher proposes competing models based on a theory. The models specify the degree of correlation between the common factors and which of the unique factors will be correlated. The different models are specified based on the initial analysis using the correlational coefficient, measurement error, and covariance. The models are specified based on researcher’s theoretical hypothesis. The competing models are tested to determine which model fits the data. Fit statistics are analyzed to determine which model best explains the relationship between observed variables and latent factors. Mulaik (1987) noted, "a goodness-of-fit test evaluates the model in terms of the fixed parameters used to specify the model, and acceptance or rejection of the model in terms of the over identifying conditions in the model" (p. 275).

In CFA more than one model might fit the data statistically, and hence finding a model that fits the data does not mean that it is the best model. The advantage of CFA as mentioned above is that different factor structures can be compared in the structural equation modeling framework. Nested models like second-order models (second-order) and bifactor models can be tested based on the a priori hypothesis. Bifactor models structures can be tested: (a) when a general factor is

hypothesized to account for the commonality of measures, (b) when there are multiple domain specific factors where each factor has a unique contribution over and above the general factor, (c) the domain specific factors are equally important as the general factor (Chen, West and Sousa, 2006). In a bifactor model, the general and domain specific factors are hypothesized to be orthogonal because the domain specific factors contribute over and above the contribution of the general factor.

Second-order or second-order models can be appropriate to use when the tests measures related domains. Second-order models are used to test the factor structure when the domain specific factors are correlated with each other and when there is a priori hypothesis that a second-order or a second-order factor can account for the relationship between the lower order factors (Chen et al., 2006). The second-order model is nested within the bifactor model, and they can be compared to check which model fits the data better. Chen et al. (2006) argue that bifactor models have several advantages over second-order models. One of the main advantages is that the role of domain specific factors can be studied independent of the general factor and the strength of the relationship between the domain specific factor and the measures can be examined which is not possible in the second-order factor model. The bifactor model also allows for easier interpretation of the data because the factor loadings of the domain specific factors are over and above the general factor. If the

general factor is the focal point of interest, this is a better model to explain the data and is more parsimonious than the second-order model.

Validity Studies in ELP Testing

There are hardly any full-fledged studies done on the validity of ELP testing. In the literature, there are two kinds of validity studies. Most of the studies address construct and how the construct is defined. There are also a few studies analyzing the predictive validity of ELP tests. The first part of this section reviews validity studies that address the use of the score from ELP tests, and second part addresses studies which look at the factor structure of different ELP tests.

Garcia, Lawton, and Diniz de Figueirido (2010), in the study on assessing young ELL students, analyzed data from the Arizona English Language Learner Assessment (AZELLA). The purpose of this study was to evaluate the relationship between AZELLA and Arizona Instrument to Measure Standards (AIMS) for ELL students. The relationship was analyzed in 3rd, 5th and 8th grade. The results from this study indicate that the reading sections from both tests are highly correlated. Third graders had the strongest correlation at 0.71. But the correlations were much lower in the higher grades. The finding from this study indicated that the tests over classifies students in higher grades. The students do not have the language support they need to function in the classroom. Garcia et al. (2010) concluded that the use of the score is a threat to validity because it fails to identify students in need of language support in the higher grades.

Another validity study done by Mahoney, Haladyna and MacSwan (2009) investigated the appropriateness of using a single language proficiency score to reclassify ELL as proficient. The data from the Stanford English Language Proficiency Test (SELP), the instrument that was used in Arizona before AZELLA, was analyzed. The findings from the study echo the views shared by Arizona Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education that it is inappropriate to use a single score to make high stakes decisions. One of the research questions addressed was if students reclassified as proficient in English by the SELP test had the necessary skills to be successful in the mainstream classroom where there is no language support. The researchers compared the performance of students reclassified by SELP with AIMS to a control group. The control group had students reclassified using multiple measures. Until 2004, the state policy allowed districts to choose the test that could be used. The control group outperformed the students compared to the SELP reclassified students. The researchers in this study concluded the SELP test over classified students as proficient, and these students did not enough language skills to function in the main stream classroom.

There are states that use multiple sources to make the determination. For example in Iowa, in addition to scoring proficient on the English Language Development Assessment (ELDA), it is recommended

that the district's exit criteria include scoring proficient on other district wide assessments and at least one or more of the following:

- Evidence of success in the regular classroom
- Sustainability of the success (one or two years)
- Lack of need for English acquisition support
- Student has been in the "Transitional" stage (one or two years)
- Teachers, other staff, and parents are in agreement
- Others (as specified by the district)

Another study analyzed the validity of ELP scores from the English Language Proficiency Test (ELPT, Bridgeman & Harvey, 1998). ELPT is a multiple-choice test to assess the ability to use English in daily interactions. ELPT consists of two subtests – listening and reading. Unlike the ELP tests used in the K-12 system, this test measured 'functional language'. This test does not have items that test grammar and usage. The listening section of this test has two types of questions. The first set was the one in which the question and the answer are read to the student and the student just has to mark the answer in the answer sheet. In the second set, the students had to listen to a report, a dialogue, or a narrative report and then answer the multiple-choice question based on what they heard. The second section was highly correlated (0.81) to the reading section. This study also correlated teacher ratings with the score on reading and listening. The results indicate that the cut scores for each of

the proficiency levels were set at a higher standard compared to teacher ratings.

Stricker, Rock and Lee (2005) conducted a CFA on LanguEdge which tests listening, speaking, reading and writing. The study addressed two questions, the factor structure of the test as well how there is a difference in the factor structure across the different groups. This study identified two distinct correlated factors – speaking and a combination of listening, reading and writing which was a different finding from the previous studies which identified three factors, a combined reading and writing factor, a speaking factor and a listening factor (Bachman, Davidson, Ryan,& Choi, 1995; Kunnan, 1995). The study also reported that there was no significant difference in the factor structure for the different groups.

Swinton and Powers' (1980) study on group differences in the factor structure of the TOEFL concluded that listening comprehension was a separate factor for all language groups in the sample. Structure, written expression, and reading comprehension loaded on one-factor for most language groups, and for those groups, vocabulary loaded on a separate factor. The conclusions from this study indicated that there are three factors, but the interpretation of the factor structure was different based on the language group. The study also reported that the level of language proficiency plays a role in interpreting the factor structures. The test was relatively easy for German students, and the factor structure showed more

highly differentiated factors, whereas for the Farsi speakers, the test was relatively difficult, and the factor structure shows the least number of factors but not interpretable.

The study on learner characteristics done by Farhady (1982) reported similar results that there are group differences in the performance of students from different cultural and educational backgrounds. He argues that the definition of English Language proficiency as used in measurement should take into account the differences between these groups. It is also pointed out in this study that test taker characteristics are not considered in the development of most English Language proficiency tests and careful consideration of this in test development process in essential to eliminate test bias.

Bachman and Palmer's (1982) study on the construct validity of communicative proficiency concludes that there is a general factor and two specific trait factors, grammatical/pragmatic competence and sociolinguistic competence. This study was done using the multitrait-multimethod approach using CFA. This study concludes that there is a second-order general factor and two first order factors. Even though grammatical and pragmatic competence was thought of as distinct factors, this study reports them as a single factor. Sociolinguistic competence is the other first order factor. All the models tested have the three correlated methods – interview, writing/ multiple-choice questions, self-rating. All

measures except the grammar multiple-choice loaded heavily on the general factor in this study.

Another recent construct validation study done on the factor structure of the internet based TOEFL test reported a second-order general English as a second language factor and four first order factors for reading, listening, speaking and writing (Sawaki, Stricker, & Oranje, 2009). This study did a CFA and tested five different models, a bifactor model, correlated four-factor model, single factor model, correlated two-factor model, and higher-order factor model. This study was done using individual items responses whereas the study by Stricker et al. (2005) was done using item parcels. This study is distinct from the other ones because it reported the model with the four first order factors and the general factor as yielding the best fit in their analysis. This is consistent with the multicomponential view of language which supports the reporting of scores in each section separately and the scores being combined to produce an overall score for language proficiency. It is also of interest that even though the second-order model yielded the best fit in this study, the speaking factor loadings were the lowest which suggests that this factor captures abilities which are not reflected in the overall language factor.

Salehi and Rezzaee (2009) report three factors for a language proficiency test that is used as part of the entrance criterion to a Ph.D. program in education at the University of Tehran called the University of Tehran English Proficiency Test. The factor analysis conducted using

principal components analysis and varimax rotation reports three factors, vocabulary, grammar and reading comprehension. The researchers did an EFA on the grammar portion of the test and reports 8 factors in one subsection with 20 multiple-choice items and six factors in the multiple-choice section with 15 items which addressed error analysis. Another study done by Salehi (2011) on the reading comprehension part of the same test reports 11 factors. This was also done using principal component analysis and a varimax rotation on the 35 item test.

Shin (2005), in his study investigating the relationship between proficiency levels and the structure of ELP tests, tested whether the structure of the language test differed based on examinee proficiency. Proficiency was measured by the grade on test called First Certificate of English (FCE). Two tests were used in this study as instruments that measure ELP, TOEFL, and the Speaking Proficiency in English Assessment Kit (SPEAK). The students were grouped into three groups – low, intermediate and high for this analysis. Different models were tested and the second-order factor model was chosen as the baseline model which is consistent with the other TOEFL study (Sawaki et al., 2009). This study concluded that there is no significant difference between the groups on factor structure which is inconsistent with the results from the Swinton and Powers (1980) study on the TOEFL test.

Chapter 3

Methods

The study examined the internal factor structure of the English language development assessment and the construct ELP as measured by ELDA. The purpose do the study was to examine the dimensionality of ELP as measured by ELDA, a language proficiency test developed based on the mandates of NCLB. Analyses were conducted to examine whether the ELP score as measured by the ELDA is a unidimensional or multidimensional construct. Alternate factor structures were also examined to determine the internal structure of the test that explained the construct of ELP.

The standard setting for the proficiency levels on the ELDA was done separately for each of the domains. This makes the results from the domains inconsistent because a student can be classified as proficient in one domain but at any of the four lower proficiency levels (pre-functional, beginning, intermediate, and advanced) in the other domains. This makes the interpretation of the results difficult and conflicting, which brings us to the issue of dimensionality. Combining the scores from the four domains into a single overall language proficiency score for placement decisions suggests ELP is regarded as more of a unidimensional construct rather than treating proficiency in each domain as requisite for classifying a student as ELP. The rules used in combining the scores for the ELDA

from each domain to an overall proficiency score is explained later on in this chapter.

Research Questions

1. Which model best represents the factor structure of the ELDA with the four language arts abilities (reading, writing, listening, and speaking)?
    i. Is the ELDA represented well by a factor structure that includes the four hypothesized factors?
    ii. How does the fit of the hypothesized four-factor model compare to a one-factor model of English language proficiency?
    iii. Is the second-order model a good fit for the data?
    iv. Can the bifactor model explain the structure of the ELDA?
2. Is there a difference between the factor structure for students in grades three to five and students in grades 9 to 12?

Instrument

ELDA is a battery of tests designed to allow schools to measure annual progress in the acquisition of English language proficiency skills among non-native English speaking students in grades kindergarten through grade 12. The battery consists of separate tests for listening, speaking, reading, and writing, at each of three grade clusters: three to five, six to eight, and 9 to 12 and a separate K to 2 Inventory. ELDA has

three forms (A, B and C). Form A was developed from the first operational test that was administered in 2005. Forms B and C were developed from field test items and linked to Form A to serve as parallel forms. The grade clusters allow the vertical linking of the test. The same items are used in adjacent grade clusters and this enabled to create the vertical linking and analysis of growth between grade clusters.

ELDA was developed by a consortium of states designed to assess the development of language proficiency as outlined by ELP standards adopted by the participating states. The development was headed by CCSSO along with the participating states in LEP-SCASS solicited proposals from different organizations and AIR was chosen for the development of the ELDA (Lara et al., 2007). ELDA was designed by combining the ELP standards from the participating states to comply with the requirements of NCLB. The standards were selected and adapted from the states that already had established ELP standards. The selection of standards to be tested on the ELDA was based on the appropriateness of the standards for each grade cluster and to fulfill the goal of English language proficiency in each grade cluster and across grade clusters (Lara et al., 2007).

The information about ELDA presented in this study came from different state and school district web sites. Most of the information came from the different informational materials put together by the Louisiana state web site and the Arkansas state web site. The information was

combined from the different sources and from the ELDA technical report (American Institutes for Research, 2005).

ELDA was designed to measure academic English as mandated by NCLB. Items were constructed from different academic content areas – English Language Arts, Mathematics, Science and Technology, and Social Studies. Items also use the context of the school environment to incorporate situational knowledge into the test. The test is designed to measure both oral and written language skills. The tests use multiple item formats to test the four domains of speaking, listening, reading and writing. There are multiple-choice items, constructed-response items and items that require the students to speak for the speaking part of the test. The constructed-response items are of two types- short constructed-response items and extended constructed-response items.

The next section provides details about the administration and scoring of each of the four tests in the ELDA. The total numbers of items for each test vary depending on the grade cluster (K to 2, three to five, six to eight, and 9 to 12). The analyses in this study were done on the three to five grade cluster and 9 to 12 grade clusters. These two grade clusters were chosen to compare whether there were differences in the factor structure for these two age groups.

Listening. All the items in the listening subtest are in the multiple-choice format. The test takers listen to different kinds of stimuli (short

passages, long passages, and conversations) and answer multiple-choice items. The content standards assessed in listening are:

1. Comprehend spoken instructions
2. Determine main idea and purpose
3. Identify important supporting details
4. Determine speaker's attitude and perspective
5. Comprehend key vocabulary and phrases
6. Draw inferences, predictions and conclusions

In the listening domain, there are 50 multiple-choice items in the three to five grade cluster and 60 items in the 9 to 12 cluster. The listening test is recorded on discs and administered to students where they listen to different types of prerecorded narrated texts and answer questions based on what they heard. The narrator reads the texts, the items, and the different options for the answers and the student is asked to record the answers in the answer booklet. The prompts are read twice but the questions and answer choices are read only once.

For example if the student is answering a question based on a short passage, either a teacher is talking with a student or two students are talking with each other. There is only one item on the test based on the short passage. The exchange is repeated twice and then the narrator reads the question and the answer options. In the longer dialogues there are two items associated with each dialogue. The number of items varies based on the grade cluster. There are 22 items for grades three to five

with long dialogues and 12 items in the 9 to 12 cluster. For the short passages the number of items changes based on the form and administration. There are four to eight items for the three to five grade cluster and 7 to 10 items in the 9 to 12 grade cluster.

Speaking. The prompts are usually graphic in nature. The students are asked to respond to multiple types of prompts so that they can show their ability to use English. These are short constructed-responses. The score ranges from zero to two. The content standards in the speaking section represent tasks of increasing complexity. The standards tested are:

1. Connect: Students are expected to have attended to the prompt which is considered a beginning level standard.
2. Tell: This is the next level where the student is given a picture prompt and asked to talk about what is represented in the picture.
3. Expand: This is considered a higher level than telling what is in the picture. The student is asked to expand on their responses.
4. Reason: This is considered as the highest level where the student is expected to go beyond expanding and is asked to draw conclusions. (Bunch, 2011)

Here is a sample item from the speaking test that assesses the first standard which is to make connections. The prompt that they will hear is below.

Some students like to listen to music in their free time. Others like to read books. Tell me in a sentence what you like to do in your free time. For example, you can talk about watching movies, listening to music, or playing with your sister or brother. Try to speak in a sentence. Tell me what you like to do in your free time.

The student gets a zero for responses like “I have free time” because the response does not address the prompt. A response like “I do everything” also does not give the student any points because the response provides not essential or specific information. The student will be given one point if the student answers in a phrase or a single word. For example, “watching movies” or in single words like ‘read’ or swim will get them 1 point. For the student to get two points for the response the student is expected to answer in a complete sentence “I play with my little sister” or in a three word phrase like “walk my dog.”

Reading. Multiple-choice items that assess reading comprehension are used in this section. The students read different kinds of material, which are of varying lengths. Multiple-choice items in this section are scored as right and wrong. One point is given for the right answer and 0 for the wrong answer. The number of questions and the standards tested are different based on grade clusters.

1. Demonstrate Pre/early reading skills
2. Comprehend key vocabulary and phrases
3. Comprehend written instructions

4. Determine main idea and purpose
5. Identify important supporting details
6. Draw inferences predictions and conclusions
7. Determine writer's attitude and perspective (only for grade clusters six to eight and 9 to 12)

Here is an example item from the 9 to 12 grade cluster. This is based on a short passage for comprehension.

> Mary and her friends Petra and David went to the mall yesterday to buy a birthday present for Petra's baby sister. They bought her a lovely toy bear.
>
> Why did Mary and her friends go to the mall?
>
> A. To see Petra's sister
> B. To get some new shoes
> C. To have lunch
> D. To buy a gift

The student has to mark the right answer in the answer booklet and gets one point if the student marked 'D,' which is the correct answer. The student will get one point for the correct answer and no points if the answer was incorrect.

Writing. The writing section has both multiple-choice items (15) and constructed-response items. The multiple-choice items are scores as right (1 point) and wrong (0 point). The constructed-response items are

scored based on a rubric and the points range from zero to four. The standards tested for writing include:

1. Planning and organizing
2. Writing a draft text
3. Revising
4. Editing

For example, in the revising part short ‘peer written’ passages are used as prompts for the multiple-choice questions. Students answer 12 multiple-choice questions where they choose the correct grammar usage or add a topic sentence or a concluding sentence. The other three multiple-choice items are to address planning and organizing where the students are given a graphic organizer and asked to choose the answer that best demonstrates the use of written English in planning and organizing content. Short and extended constructed-response items are used to address the other standards. In the three to five grade cluster there are three short constructed-response items and one extended-response item whereas in the 9 to 12 there are four short constructed-response items and one extended-response item.

Performance Levels

The raw scores from each of the domains are converted to scale scores and proficiency levels for each of the domains are reported. ELDA reports five proficiency levels for each domain and an overall proficiency level. The five proficiency levels are

1. Pre-functional
2. Beginning
3. Intermediate
4. Advanced
5. Full English Proficiency

The cut scores for the proficiency levels were established by a bookmark standard to setting process. Detailed information about this is provided in the technical report (Bunch, 2006). The test was vertically scaled and hence the cut points are different for each domain and grade cluster. For each domain, a proficiency level is calculated for each student. A student who takes all four part of the test is assigned a proficiency level for speaking, a level for writing, a level for reading, and a level for listening. The proficiency levels were set based on the recommendations from the articulation committee and the technical advisory group (Bunch, 2006). In addition to the proficiency levels reported for each domain, the test also reports a comprehension level which is a combination of proficiency levels the students received on listening and reading. Table 1 shows how listening level is combined with the reading level from the test to create the comprehension level. A production level score is computed which is a combination of speaking and writing. Table 2 displays this conversion. The composite level or the overall level is calculated from combining the production and the

comprehension levels. Table 3 shows how the overall composite level is calculated by combining the production and comprehension levels.

Table 1

Rules for Computing Comprehension Level From Listening and Reading Level

Reading	Listening				
	1	2	3	4	5
1	1	1	1	2	2
2	2	2	2	2	3
3	2	3	3	3	3
4	3	3	4	4	4
5	3	3	4	5	5

Note. (1 = pre-functional, 2= beginner, 3 = intermediate, 4 = advanced, 5 = fully English proficient)

Table 2

Rules for Computing Production Level from Speaking and Writing Level

Writing	Speaking				
	1	2	3	4	5
1	1	1	1	2	2
2	2	2	2	2	3
3	2	3	3	3	3
4	3	3	4	4	4
5	3	3	4	5	5

Note. (1 = pre-functional, 2= beginner, 3 = intermediate, 4 = advanced, 5 = fully English proficient)

Table 3

Rules for Composite Level from Comprehension and Production

Production	Comprehension				
	1	2	3	4	5
1	1	1	2	2	3
2	1	2	2	3	3
3	2	2	3	3	4
4	2	3	3	4	4
5	3	3	4	4	5

Note. (1 = pre-functional, 2= beginner, 3 = intermediate, 4 = advanced, 5 = fully English proficient)

Psychometric Properties of the Test

The information put together in this part came from the technical report (American Institutes for Research, 2005) for the test. Test difficulties range from p=0.54 for writing in grade cluster six to eight to p=0.81 for speaking in grade clusters three to five and 9 to 12. Test difficulties are comparable across grade clusters in each skill domain. The reliability estimates for each of the forms and domains were calculated using Cronbach's alpha. The reliability coefficients indicate consistently high reliability for the test for the three different forms of the test and for the four domains. The reliability coefficients range from 0.76 to 0.95. The reliability is relatively lower for the writing test compared to the other three domains. The Table 4 provides the reliability coefficients for the three

forms (A, B, and C) and the four domains (listening, speaking, reading, and writing).

Table 4

Reliability Coefficients(Cronbach's alpha)

Domain	Form	Grade cluster three to five	Grade cluster 6 to 8	Grade cluster 9 to 12
Listening	A	0.91	0.93	0.94
	B	0.92	0.92	0.95
	C	0.92	0.93	0.94
Reading	A	0.93	0.93	0.95
	B	0.93	0.93	0.94
	C	0.93	0.94	0.95
Speaking	A	0.89	0.94	0.90
	B	0.90	0.93	0.88
	C	0.88	0.93	0.92
Writing	A	0.76	0.85	0.84
	B	0.79	0.85	0.84
	C	0.82	0.84	0.86
	D	0.79	0.84	0.87

Data

ELDA produces a score for each of the domains separately and a composite score which combines the scores across each of the domains.

As described in the previous section the domains are unequally weighted when the overall proficiency level is determined. The data were received from CCSSO. The data included item level responses for each item for all grades (kindergarten to 12th) for each of the domains. The data set did not have any identifying information about the students. The data used for this study was from grade clusters three to five and 9 to 12. The two grade clusters were chosen to examine the similarities and differences in the two age groups.

Items were combined together in each domain based on the standard being assessed by the item. The total raw score computed for each standard within each domain was used for this analysis. The raw score from each of the items for the content standard were added together to get the total score for the item parcels in each domain. This was included in the data set for each standard in each of the domains. There were seven different content standards assessed in reading, six in listening, four in writing and four in speaking which added up to 21 item parcels. The content standards tested are listed under the description of the instrument in the first part of the chapter. There were 4,577 observations in the three to five sample and 2,330 observations in the 9 to 12 sample.

Participants

The participants in this study were ELL students from different states in the US. The data came from the administration of the ELDA

(Form A) in 2008. The sample used in this study consisted of students in grades three to five and 9 to 12. There were students from all proficiency levels represented in the sample for each of the domains. For the three to five sample, 50 students were omitted because they did not complete all four parts of the ELDA. For the 9 to 12 sample, 47 students were omitted.

Analyses

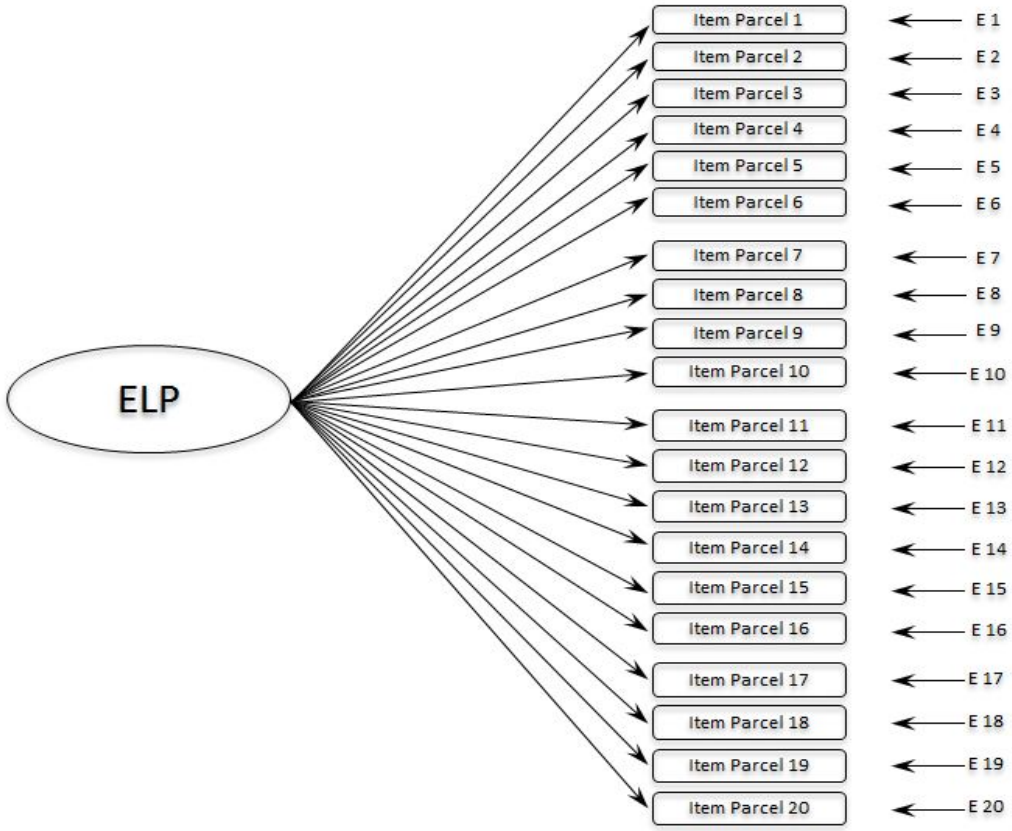
The initial data analysis was conducted using Statistical Package for the Social Sciences (SPSS 20). Only students in the sample with a score in all domains will be included in the analyses. Means, standard deviations, and product-moment correlation indices were computed for the item parcels in each domain to describe the data. Frequencies of proficiency levels for each domain and the overall levels were computed to explain the distribution of data.

The research questions were answered by conducting confirmatory factor analyses using the maximum likelihood parameter estimates with standard errors and a mean-adjusted chi-square test statistic (MLM) that are robust to non-normality. Different models were compared to understand the factor structure of the ELDA. Figures 1 to 5 illustrates the different models tested. Item parcels one to six are the reading items, 7 to 10 is writing, 11 to 16 is speaking, and 17 to 20 is speaking. Five models were compared to test the hypotheses about the factor structure. They were: single factor model, correlated two-factor model, correlated four-

factor model, bifactor model and the higher order model. The models are described below.

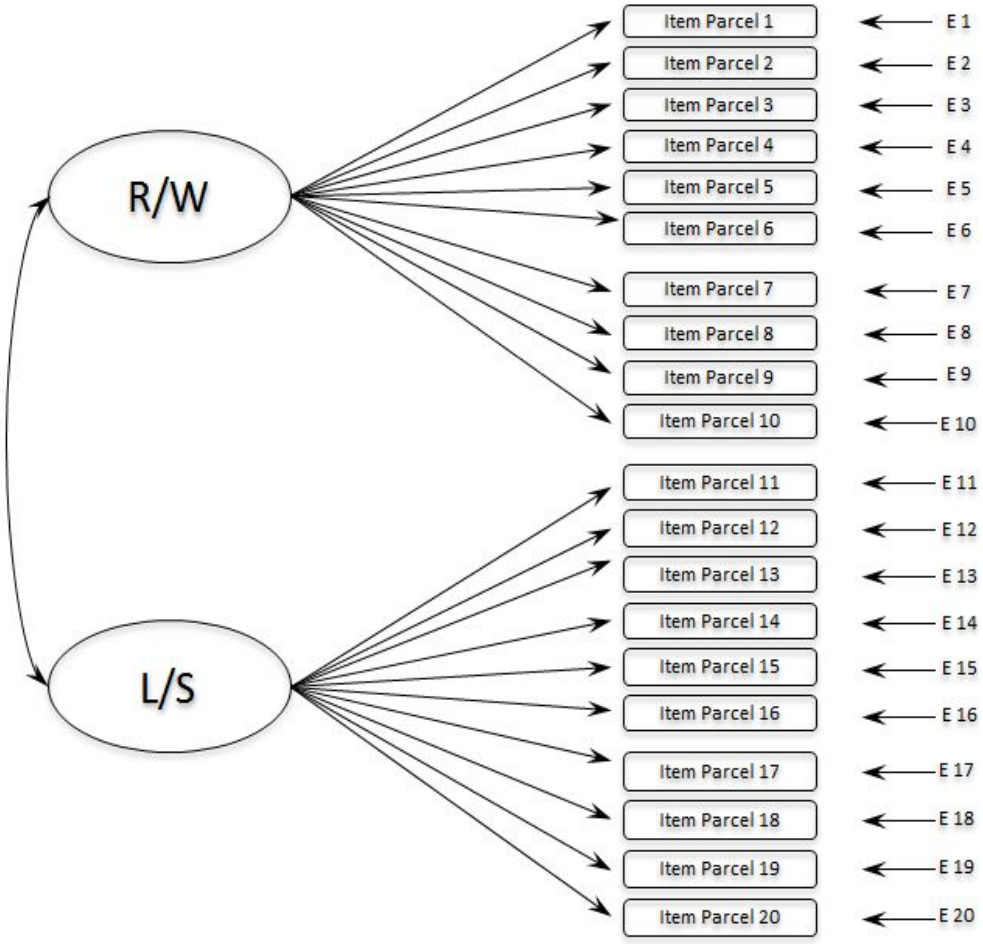
- The Single Factor Model: This model tests the hypothesis that there is only one-factor that represents all item parcels in the four domains of listening, speaking, reading and writing. This model assumes that the measures from the four domains are interchangeable. In the three to five cluster the 20 different items parcels were treated as measures with one underlying factor, second language proficiency. There were 21 item parcels that were used as measures in the 9 to 12 grade cluster. Figure 1 illustrates this model for the three to five grade cluster with 20 item parcels.

Figure 1. Single factor model



- The Two-factor Model: This model tests the hypothesis that there are two correlated factors. The first factor is a combination of listening and speaking. The second factor is a combination of reading and writing. Figure 2 illustrates this model for the three to five grade cluster with 20 item parcels. The six listening measures and the four speaking measures were allowed to load on the language factor. The six reading measures and the four writing measures were allowed to load on the literacy factor. In the 9 to 12 cluster there were 7 reading measures that were used.

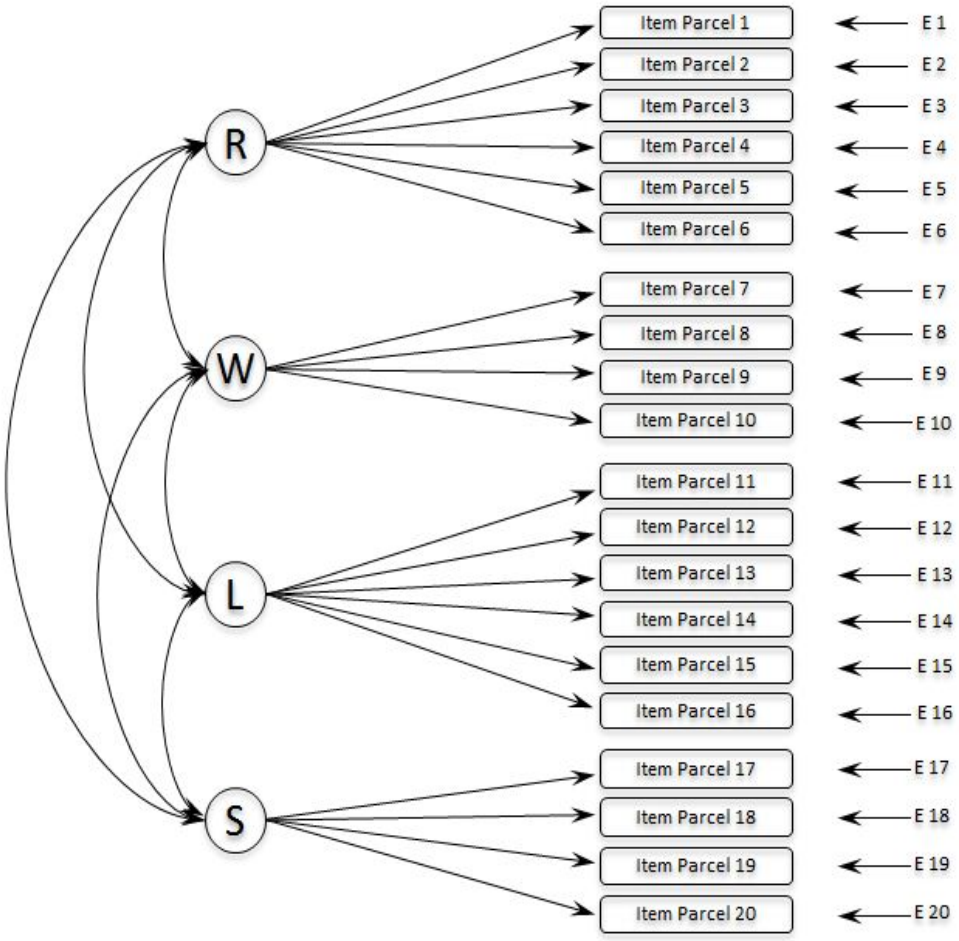
Figure 2. The two-factor model



- The Correlated Four-factor Model: This model tests the hypothesis that there are four distinct correlated factors listening, speaking, reading, and writing. The six item parcels which were reading measures underlying the first factor, the four writing item parcels loaded on the writing factor, the six listening item parcels loaded on the listening factor and the four items parcels that measured speaking loaded on the reading factor. Figure 3 illustrates this

model for the three to five grade cluster. The 9 to 12 grade cluster has seven items parcels that contribute to the reading factor.

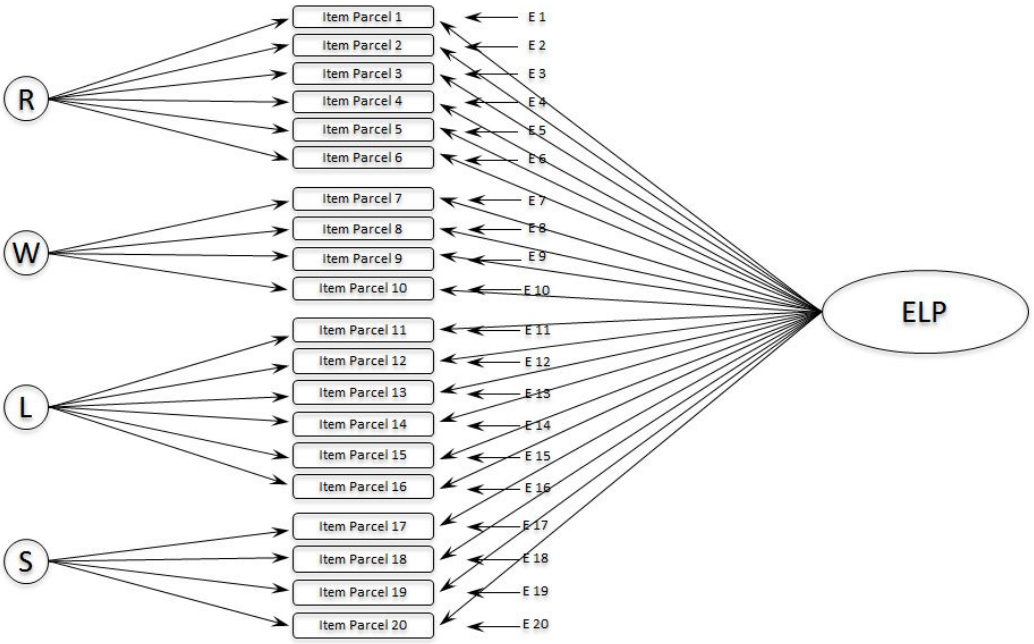
Figure 3. Correlated four-factor model



- The Bifactor Model: This model tests the hypothesis that there is a general language factor as well as the four other factors of listening, speaking, reading, and writing that explain the factor structure of ELP as measured by ELDA. In this model the second language general factor is represented by each of the measures in

the model. The general language factor in this model explains the commonality of the item parcels. The domain specific factors in this model are hypothesized to account for the unique contribution of the groups of item parcels over and above the general factor. The bifactor model is recommended by Chen, Sousa and West (2006) as a better model to explain the relationship between highly related domains. Figure 4 illustrates this model for the three to five grade cluster.

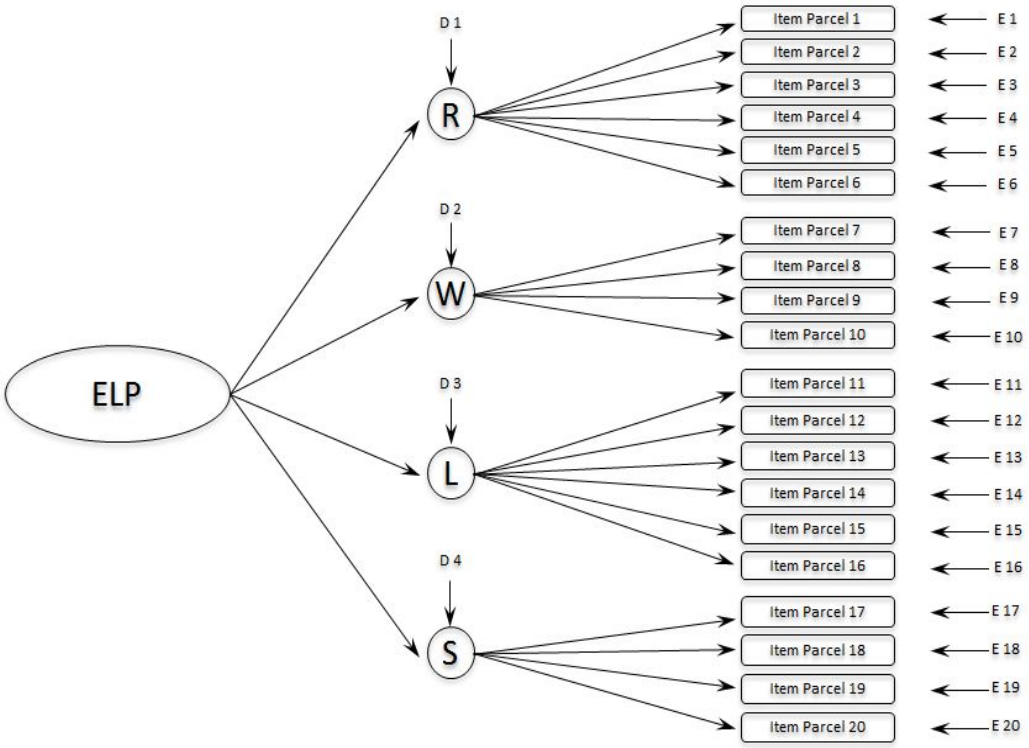
Figure 4. Bifactor model



- A Second-order Factor Model: This model hypothesizes a general language factor that accounts for the relationship between the four domains of speaking, listening, reading and writing. In this model

the overall language factor is represented by each of the lower order factors. Figure 5 represents this model for the three to five grade cluster.

Figure 5. Second-order model



Model fit was evaluated based on different indices, the chi-square statistic, root mean squared error of approximation (RMSEA, ≤0.06 for good fit and <0.10 for adequate fit), standardized root mean residual (SRMR, <0.08), and comparative fit index (CFI, >0.90-0.95) (Hu & Bentler, 1999). The different models were compared to understand the factor structure of the ELDA. Scaled Satorra-Bentler chi- square difference tests

were used to evaluate the models nested within the bifactor model (i.e., the second-order model and one-factor model).

The second research question examines whether there is a difference in the factor structure underlying the ELDA for different grade clusters. The CFA analyses conducted on the three to five grade cluster was repeated for the 9 to 12 grade cluster to check for similarities and differences in the measures in the two grade clusters.

Chapter 4

Results

Sample

In the sample, the students were clustered in the three proficiency levels (beginning, intermediate, and advanced) based on the overall proficiency level reported. It should also be noted that less than 1% of students were classified as proficient based on the overall proficiency level. Table 5 shows the percentages of students in the five different proficiency levels (level 1 = pre-functional, level 2 = beginning, level 3= intermediate, level 4 = advanced, level 5= fully English proficient) for each domain. There was some variability in the levels assigned for the different domains. The greatest number of students was classified as fully English proficient on the speaking test, and writing scores yielded the fewest number of students considered proficient.

Table 5

Proficiency Level Distribution (N (3 to 5) =4,577 and N (9 to 12) =2,330)

Subject	Grade	Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)	Level 5 (%)
Reading	3 to 5	15.34	25.76	13.24	29.34	16.32
	9 to 12	25.02	19.06	20.77	24.08	11.07
Writing	3 to 5	4.70	31.79	34.04	29.06	0.42
	9 to 12	12.40	29.79	28.80	24.89	4.12
Listening	3 to 5	4.85	9.70	23.40	27.90	34.15
	9 to 12	12.96	9.36	16.18	28.11	33.39
Speaking	3 to 5	2.03	1.16	8.54	29.78	58.49
	9 to 12	8.63	4.81	10.64	13.78	62.15
Overall	3 to 5	10.09	29.67	31.46	28.47	0.31
	9 to 12	18.80	24.29	33.35	20.94	2.62

The raw scores ranged from 0 to 50 for reading and listening in the three to five grade cluster, and the range went from 0 to 60 in the 9 to 12 cluster. The writing scores for the three to five cluster ranged from 0 to 25 whereas in the 9 to 12 cluster, the scores ranged from 2 to 34 points. The speaking raw score ranges for both clusters were the same. The range was from 0 to 32 for both clusters. The students scored the highest in speaking and lowest in writing. The percent of points earned for each of the domains for both grade clusters is shown in Table 6.

Table 6

Average Percent of Points (N (3 to 5) =4,577 and N (9 to 12) =2,330)

Subject	Grade	Average percent points	Total points possible
Reading	3 to 5	68.6	50
	9 to 12	62.1	60
Writing	3 to 5	59.4	28
	9 to 12	64.4	34
Listening	3 to 5	77.7	50
	9 to 12	74.0	60
Speaking	3 to 5	87.2	32
	9 to 12	83.8	32

The items were combined based on the standard being tested into item parcels in each of the domains. There were six item parcels for reading in the three to five cluster and seven in the 9 to 12 cluster. All other item parcels were the same number for both clusters; however, the maximum number of points possible was different for both grade clusters because of the difference in the number of items in each item parcel. The percent of points earned was highest in speaking and lowest in reading. Table 7 illustrates these differences in the total possible points and displays the percent of points earned in each of the item parcels for both grade clusters three to five and 9 to 12 in the four domains.

Table 7

Average percent points (N (3 to 5) =4,577 and N (9 to 12) =2,330)

Domain	Measures	Grades	Average percent points	Total points possible
Reading	Demonstrate pre-early reading skills	3 to 5	86.6	12
		9 to 12	71.3	12
	Comprehend key vocabulary and phrases	3 to 5	56.3	7
		9 to 12	57.3	6
	Comprehend written instructions	3 to 5	68.8	6
		9 to 12	69.3	8
	Determine main idea and purpose	3 to 5	64.5	3
		9 to 12	67.5	4
	Identify important supporting details	3 to 5	68.1	14
		9 to 12	59.4	20
	Draw inferences, predictions and conclusions	3 to 5	54.8	8
		9 to 12	50.4	8
	Determine writer's attitude and perspective	3 to 5	--	--
		9 to 12	54.9	2
Writing	Planning and organizing	3 to 5	56.0	6
		9 to 12	59.0	6
	Writing a draft text	3 to 5	61.5	13
		9 to 12	65.6	16
	Revising	3 to 5	49.7	3
		9 to 12	51.3	3
	Editing	3 to 5	64.9	3
		9 to 12	70.4	9
Listening	Determine main idea and purpose	3 to 5	81.6	7
		9 to 12	72.4	10
	Comprehend spoken instructions	3 to 5	82.2	12
		9 to 12	78.9	7
	Identify important supporting details	3 to 5	69.7	14
		9 to 12	67.7	14
	Determine speaker's attitude and perspective	3 to 5	68.3	4
		9 to 12	70.1	7
	Comprehend key vocabulary and phrases	3 to 5	81.6	8
		9 to 12	83.0	11
	Draw inferences, predictions and conclusions	3 to 5	84.8	5
		9 to 12	73.1	11
Speaking	Connect	3 to 5	93.3	8
		9 to 12	88.8	8
	Tell	3 to 5	94.7	8
		9 to 12	88.4	8
	Expand	3 to 5	87.3	8
		9 to 12	85.6	8
	Reason	3 to 5	73.4	8
		9 to 12	72.3	8

Correlations

The correlations between raw scores from the four domains were statistically significant in both three to five and the 9 to12 grade clusters. In both groups, the speaking domain had the lowest correlations with the other domains. The reading raw scores were highly correlated with writing, 0.76 in the three to five cluster and 0.80 in the 9 to12 cluster. The correlations between all four domains in the 9 to 12 grade cluster were higher compared to the three to five grade cluster. Table 8 shows the correlations between the raw scores for the four domains.

Table 8

Correlations among the Total Raw Scores

Subject	Reading	Writing	Listening	Speaking
Reading	--	0.80	0.79	0.58
Writing	0.76	--	0.76	0.63
Listening	0.72	0.65	--	0.72
Speaking	0.50	0.53	0.62	--

Note. The correlations of the three to five grade cluster is below the diagonal and the 9 to 12 cluster is above the diagonal.

The correlations between item parcels in each of the four domains revealed slightly different trends than the overall correlations noted above. There were a total of twenty item parcels in the three to five grade cluster and twenty-one in the 9 to 12 cluster. Table 8 shows the correlations among the item parcels used in the analyses. The correlations above the diagonal are for the 9 to 12 cluster and the ones below are for the three to five cluster.

Table 9

Correlations among the Measures in Reading, Writing, Speaking and Listening

Measures	R_Erd	R_Voc	R_Instr	R_MIdea	R_SIdea	R_Inf	R_att	W_Plan	W_CR	W_Rev	W_Edit	L_Dir	L_MI	L_Det	L_Att	L_Voc	L_Inf	S_Con	S_Tell	S_Exp	S_Rea
R_Erd	NA	0.57	0.64	0.57	0.66	0.61	0.49	0.43	0.67	0.42	0.60	0.65	0.67	0.70	0.61	0.64	0.66	0.51	0.53	0.55	0.61
R_Voc	0.45	NA	0.56	0.55	0.72	0.67	0.48	0.47	0.55	0.40	0.52	0.44	0.53	0.58	0.53	0.42	0.47	0.32	0.36	0.36	0.40
R-Instr	0.53	0.48	NA	0.56	0.66	0.54	0.45	0.48	0.63	0.42	0.56	0.53	0.58	0.61	0.53	0.56	0.56	0.45	0.47	0.49	0.49
R_MIdea	0.42	0.53	0.46	NA	0.68	0.57	0.50	0.44	0.53	0.36	0.51	0.50	0.55	0.56	0.50	0.48	0.51	0.37	0.39	0.41	0.44
R_SIdea	0.57	0.68	0.60	0.61	NA	0.73	0.58	0.55	0.63	0.46	0.60	0.54	0.63	0.69	0.64	0.53	0.60	0.41	0.44	0.46	0.50
R_Inf	0.46	0.64	0.50	0.57	0.71	NA	0.53	0.47	0.56	0.40	0.51	0.48	0.56	0.60	0.55	0.46	0.53	0.37	0.40	0.40	0.46
R_att	--	--	--	--	--	--	NA	0.34	0.44	0.33	0.40	0.41	0.45	0.51	0.45	0.40	0.47	0.32	0.32	0.33	0.40
W_Plan	0.40	0.51	0.46	0.46	0.55	0.52	--	NA	0.48	0.43	0.54	0.36	0.42	0.48	0.43	0.37	0.39	0.26	0.30	0.29	0.29
W_CR	0.54	0.45	0.48	0.43	0.56	0.47	--	0.42	NA	0.43	0.60	0.63	0.65	0.66	0.59	0.66	0.63	0.63	0.65	0.67	0.65
W_Rev	0.29	0.37	0.33	0.33	0.42	0.38	--	0.37	0.32	NA	0.48	0.37	0.40	0.43	0.39	0.37	0.40	0.24	0.27	0.28	0.30
W_Edit	0.33	0.38	0.35	0.36	0.43	0.40	--	0.39	0.37	0.36	NA	0.51	0.56	0.58	0.52	0.52	0.55	0.42	0.44	0.44	0.46
L_Dir	0.50	0.44	0.46	0.42	0.53	0.47	--	0.41	0.51	0.30	0.32	NA	0.70	0.69	0.62	0.77	0.75	0.60	0.61	0.60	0.63
L_MI	0.48	0.44	0.43	0.43	0.51	0.45	--	0.39	0.49	0.28	0.30	0.65	NA	0.78	0.70	0.73	0.75	0.56	0.56	0.57	0.61
L_Det	0.48	0.53	0.48	0.48	0.59	0.53	--	0.47	0.47	0.33	0.33	0.66	0.63	NA	0.75	0.70	0.75	0.53	0.55	0.55	0.62
L_Att	0.33	0.36	0.33	0.32	0.39	0.37	--	0.33	0.34	0.22	0.23	0.45	0.43	0.49	NA	0.63	0.67	0.48	0.48	0.49	0.54
L_Voc	0.48	0.45	0.44	0.43	0.53	0.47	--	0.42	0.50	0.31	0.33	0.67	0.63	0.64	0.44	NA	0.76	0.67	0.66	0.67	0.68
L_Inf	0.45	0.42	0.41	0.39	0.49	0.44	--	0.37	0.46	0.26	0.30	0.62	0.59	0.60	0.42	0.61	NA	0.57	0.57	0.58	0.63
S_Con	0.38	0.26	0.28	0.26	0.33	0.28	--	0.24	0.49	0.18	0.22	0.45	0.43	0.36	0.26	0.44	0.41	NA	0.84	0.83	0.77
S_Tell	0.40	0.24	0.27	0.24	0.31	0.26	--	0.23	0.48	0.16	0.22	0.45	0.44	0.35	0.26	0.44	0.42	0.70	NA	0.86	0.77
S_Exp	0.40	0.32	0.33	0.31	0.38	0.33	--	0.29	0.49	0.19	0.25	0.49	0.47	0.42	0.32	0.48	0.48	0.64	0.69	NA	0.81
S_Reas	0.39	0.36	0.35	0.35	0.42	0.39	--	0.32	0.45	0.23	0.26	0.48	0.47	0.46	0.33	0.48	0.47	0.54	0.54	0.69	NA

Note. Measures with acronym in parantheses. Reading measures have R as their first letter in the name of the measure, W for writing, L for listening and Sp for speaking. Demonstrate early reading skills (R_PreRead), comprehend key vocabulary and phrases (R_Vocab), comprehend written instructions (R-Instr), determine main idea and purpose (R_MainIdea), identify important supporting details (R_SupIdea), draw inferences predictions and conclusions (R_InfComp), determine writers attitude and perspective (R_Att), Planning and organizing (W_Plan), writing a draft text (W_CR), revising (W_Rev), editing (W_Edit), determine main idea and purpose (L_MI), Comprehend spoken instructions (L_MI), identify important supporting details (L_Det), determine speakers attitude and perspective (L_SpAtt), comprehend key vocabulary and phrases (L_Vocab), draw inferences predictions and conclusions (L_Inf), Connect (Sp_Con), tell (Sp_Tell), Expand (S_Exp), reason (Sp_Reason)

Confirmatory Factor Analyses

Five different confirmatory factor analyses were conducted to answer the first research question in this study, which was to examine the internal factor structure of the ELDA. The analyses were repeated for the 9 to 12 grade cluster to answer the second research question in the study, which addressed whether there is a difference in the factor structure between the three to five grade cluster and the 9 to 12 grade cluster.

One-factor model. The model hypothesized a single language factor that could account for all of the covariance between the measures. The scale of the factor was defined by setting the factor loading of one measure to be 1. The data did not fit this model well, χ^2 (189) = 8,165.75, $p < 0.01$, RMSEA = 0.14, SRMR = 0.08, and CFI = 0.78 for the 9 to 12 cluster. The three to five grade cluster showed similar results, χ2 (170) = 8,939.49, $p < 0.01$, RMSEA = 0.12, SRMR = 0.07, and CFI = 0.79.

Correlated two-factor model. The two-factor model hypothesized there are two-factors, a combined factor with listening and speaking and another combined factor consisting of reading and writing. The measures were allowed to load on only two-factors, language (speaking and listening measures) and literacy (reading and writing). The model was identified by setting the factor loading of one measure in each of the factor to be 1. The model fit indices indicate that the data did not fit the model well, χ^2 (169) = 5,562.32, $p < 0.01$, RMSEA = 0.12, SRMR = 0.07 and CFI = 0.84

for the three to five cluster. The 9 to 12 grade cluster showed similar results, χ2 (164) = 6002.99, $p < 0.01$, RMSEA = 0.08, SRMR = 0.06 and CFI = 0.87. The parameter estimates were significant and each of the measures loaded highly on each of the factors as expected in this model. The correlations between the two-factors were much higher in the 9 to12 cluster compared to the three to five grade cluster. The correlations between the factors were 0.76 and 0.83 in the three to five and 9 to 12 grade clusters respectively.

Correlated four-factor model. The four-factor model reflects the hypothesis of four correlated domains underlying the parcels: reading, writing, speaking, and listening. Measures were defined to load only on their intended factors. The model was identified by setting the factor loading of one measure on each factor be one. The test of model fit indicates good fit for this model, χ^2 (183) = 2,297.84, $p < 0.01$, RMSEA = 0.07, SRMR = 0.05 and CFI = 0.94 for the 9 to 12 cluster. The three to five grade cluster showed similar results, χ2 (164) = 2134.62, $p < 0.01$, RMSEA = 0.05, SRMR = 0.04 and CFI = 0.95. The parameter estimates were significant and each of the measures loaded highly on each of the factors as expected in this model.

The correlations between the factors were very high. For the three to five cluster, the correlation between the reading factor and writing factor was 0.94, writing and listening was 0.81 and reading with listening was 0.79. The speaking factor was the least correlated with the other factors.

The correlations with the speaking factor were 0.53, 0.65, and 0.70 with reading, writing, and listening respectively. For the 9 to 12 cluster the correlations between the reading and writing factor was 0.91, reading and listening was 0.84 and writing with listening was 0.87. The speaking factor was the least correlated with the other factors in the 9 to 12 grade cluster but the correlations were higher compared to the three to five cluster. The speaking factor was the least correlated with the reading factor and the most correlated with the listening factor. The correlations between the speaking factor were 0.60, 0.73 and 0.75 with reading, writing and listening respectively.

Second-order model. The second-order model hypothesized five factors, the four domain specific factors and a higher-order language factor which accounts for the relationship among the domain specific factors. In other words, the relationship among the first order factors could be explained by the general second language factor in this model. The model was defined by constraining each measure to have a zero loading on the first order factor that it was supposed to measure, and it was not allowed to load on any other factors. One loading from each domain was set to 1. The model fit indices indicate adequate fit for both grade clusters, χ^2 (185) = 2,533.77, $p < 0.01$, RMSEA = 0.07, SRMR = 0.06 and CFI = 0.93 for the 9 to 12 cluster and χ^2 (166) = 2,590.37, $p < 0.01$, RMSEA = 0.06, SRMR = 0.05 and CFI = 0.94 for the three to five cluster.

The parameter estimates were significant, and the loadings for the general factor were higher than the loadings for each of the first order factors.

The correlations between the factors were very high in both the three to five and 9 to 12 grade clusters. The general factor had almost perfect correlation with writing (0.99) in the three to five grade cluster and had a very high correlation (0.97) in the 9 to 12 grade cluster. The correlations between the other factors to the general factor were 0.91, 0.87, and 0.67 for reading, listening and speaking respectively for the three to five grade cluster. For the 9 to 12 grade cluster the correlations with the general factor were 0.91, 0.92, and 0.75 for reading, listening and speaking respectively.

The bifactor model. The test of the bifactor model hypothesized that the structure of the ELDA could be explained by five factors, the overall general second language factor and the four domain specific factors of speaking, reading, writing, and listening. The model was defined as (a) each item parcel was allowed to load on the factor that it was meant to measure and on the general factor, (b) The factor covariances were set to zero. The scale of each factor was set by fixing the factor loading of one measure in each of the domain specific factors to one. The general language factor also had one the reading measures (identifying important supporting details) set to one. The fit indices showed good fit for this model, χ^2 (168) = 1,700.03, $p < 0.01$; RMSEA = 0.06; SRMR = 0.04, and CFI = 0.96 for the 9 to 12 grade cluster. The

indices were similar for the 3 to 5 grade cluster, χ^2 (150) = 1,906.36, p < 0.01; RMSEA = 0.05; SRMR = 0.04, and CFI = 0.96. Table 10 gives the fit indices for all models tested.

The factor loadings for all measures except two measures in the three to five grade cluster were significant (p < 0.01). The reading early literacy skills measure had a negative factor loading, and it was not significant (p = 0.07), and the writing constructed-response measure also had a negative factor loading (p = 0.10) in the three to five cluster. In the 9 to 12 cluster all measures had positive loadings and were significant (p < 0.01). The general overall second language factor had the highest loadings which indicate that the model fit the data well. The loadings ranged from 0.46 to 0.80 on the general second language factor for the three to five cluster and from 0.50 to 0.85 in the 9 to 12 cluster. These results support that the four domains of reading, writing, listening, and reading account for covariation among the item parcels over and above the general second language factor. Table 11 lists the standardized factor loadings for the bifactor model.

Table 10

Summary of Fit Statistics for the Bifactor Model Compared to the Other Models

	Grade	χ^2	df	S-B Diff χ^2	S-B Diff χ^2 df	CFI	RMSEA	SRMR
Single factor	3 to 5	8,939.49	170	5807.18	20	0.79	0.12	0.07
	9 to 12	8,165.75	189	5738.49	21	0.78	0.14	0.08
Second-order	3 to 5	2,590.37	166	654.87	16	0.94	0.06	0.05
	9 to 12	2,533.77	185	801.08	17	0.93	0.07	0.06
Bifactor	3 to 5	1,906.36	150			0.96	0.05	0.04
	9 to 12	1,700.03	168			0.96	0.06	0.04

Table 11

Standardized Factor Loadings from the Bifactor, Second-order, and the Four-factor Model for 9 to12 Grade Cluster

	Bifactor model						Second-order model				Four-factor model			
Measures	ESL factor	Read	Write	Listen	Speak	Higher Order factor	Read	Write	Listen	Speak	Read	Write	Listen	Speak
R_Erd	0.82	0.07					0.80				0.80			
R_Voc	0.65	0.46					0.77				0.77			
R-Instr	0.74	0.16					0.76				0.76			
R_MIdea	0.66	0.32					0.74				0.74			
R_SIdea	0.78	0.47					0.88				0.89			
R_Inf	0.69	0.43					0.79				0.79			
R_att	0.56	0.30					0.64				0.64			
W_Plan	0.55		0.44					0.60				0.60		
W_CR	0.82		0.06					0.84				0.84		
W_Rev	0.50		0.35					0.55				0.55		
W_Edit	0.70		0.35					0.74				0.74		
L_Dir	0.75			0.39					0.82				0.83	
L_MI	0.82			0.27					0.87				0.87	
L_Det	0.85			0.18					0.88				0.88	
L_Att	0.77			0.17					0.80				0.79	
L_Voc	0.77			0.39					0.84				0.85	
L_Inf	0.79			0.38					0.87				0.86	
S_Con	0.65				0.63					0.90				0.90
S_Tell	0.67				0.63					0.92				0.92
S_Exp	0.68				0.64					0.93				0.93
S_Reas	0.73				0.48					0.86				0.86

Model Comparisons

As mentioned above the three best models that fit the data were the bifactor model, the four-factor model, and the second-order model. Table 9 shows the scaled chi-square difference test from each of the comparisons of nested models. The bifactor model was compared to the two other nested models (the one-factor and the second-order), and it was determined that the bifactor model was the best model based on the Satorra-Bentler chi-square difference test. The bifactor model was considered as the baseline model, and all the other models were compared to this model.

Bifactor model versus single factor model. The single factor model which hypothesized only one overall language factor did not fit the data well based on fit indices. However, this model was compared to the bifactor model to compare the fit. The results from the S-B scaled chi-square difference test indicate that the bifactor is model is a better fit for the data ($\chi 2_{\text{S- B difference}}$ (20) = 5807.18, $p < 0.01$) in the three to five cluster. The 9 to12 cluster indicated similar results, ($\chi 2_{\text{S- B difference}}$ (21) = 5738.49, $p < 0.01$). The chi-square difference test was significant which means that the less constrained bifactor model fit the data better than the more parsimonious single factor model fit.

Bifactor model versus second-order factor model. The bifactor model was compared to the higher/second-order factor model. Even though the interpretations of the two models are similar, these two models

are mathematically different. The second-order model is nested within the less restricted bifactor model and hence the two models were compared using a chi-square difference test. The chi-square difference indicated that the fit of the bifactor model was significantly better than the second-order factor model ($\chi^2_{\text{S- B difference}}$ (16) = 654.87, $p < 0.01$) for the 3 to 5 grade cluster. This was true for the 9 to 12 grade cluster as well ($\chi^2_{\text{S- B difference}}$ (17) = 801.06, $p < 0.01$).

Second-order model versus the four-factor model. The second-order model was compared to the four-factor model. The chi-square difference test was significant ($\chi^2_{\text{S- B difference}}$ (2) = 380.67, $p < 0.01$) for the 3 to 5 grade cluster, as well as for the 9 to 12 grade cluster ($\chi^2_{\text{S- B difference}}$ (2) = 195.08, $p < 0.01$). The scaled chi-square difference test indicates that the second-order model was better to explain the data.

Four-factor model versus the one-factor model. The four-factor model was compared to the one-factor model. The one-factor model hypothesized that the tests measure the unidimensional construct, ELP which assumes there is no distinction between the four domains. The chi-square difference test was significant ($\chi^2_{\text{S- B difference}}$ (6) = 4059.25, $p < 0.01$) for the 3 to 5 grade cluster, as well as for the 9 to 12 grade cluster ($\chi^2_{\text{S- B difference}}$ (6) = 4147.42, $p < 0.01$) indicating that the less constrained four-factor model is better to explain the data.

Two-factor model versus the one-factor model. The two-factor model was also compared to the one-factor model to check the fit of these

models. The chi-square difference test was significant ($\chi^2_{\text{S- B difference}}$ (1) = 2377.44, $p < 0.01$) for the 3 to 5 grade cluster, as well as for the 9 to 12 grade cluster ($\chi^2_{\text{S- B difference}}$ (1) = 1864.47, $p < 0.01$) indicating that the less constrained two-factor model fits the data better compared to the one-factor model.

Chapter 5

Discussion and Conclusions

Three of the proposed models fit the data well: the correlated four-factor model, the second-order model, and the bifactor model. Based on the RMSEA, CFI, and SRMR, the bifactor model seems to be the best fit of the three models for both grade clusters three to five and 9 to 12. The scaled Satorra-Bentler chi-square difference test also indicated that the bifactor model best represents the structure of the ELDA. This finding has implications in the scoring and reporting of proficiency levels which is addressed later in this chapter. The second research question in this study was to examine whether there is a difference in factor structures between the younger students who are in elementary grades (three to five) and those in the high school grades (9 to 12). Similar results for each model and for the model comparisons were obtained for these two age groups. The indices from the 9 to 12 cluster are slightly better than the three to five grade cluster in all models. However, there were differences in the factor loadings for each of the models between the two grade clusters.

Bifactor Model

The results from testing the bifactor model indicate that each domain (listening, speaking, reading, and writing) has a unique contribution to the construct of ELP over and above the general factor that is measured by the instrument. The factor loadings for this model were significant for the general factor as well as for the each of the domain

specific factors in the 9 to 12 grade cluster. In the three to five cluster in the bifactor model, however, there were two standardized parameter estimates that were low, negative, and not significant in their contribution to the domain specific factors. The first measure that yielded a negative coefficient was early pre-reading literacy skills within the reading domain and the second one was the constructed-response measure in the writing domain. In the 9 to 12 grade cluster, the standardized coefficients were low even though they were significant for these two measures.

The factor loadings for measures loading on the speaking domain were the highest in this model. The standardized loadings for the four speaking measures were high on the domain specific loadings (0.45 – 0.67 for three to five and 0.48-0.64 for 9 to 12) as well as on the general factor (0.49-58 for three to five and 0.65-0.73) in both the three to five and the 9 to 12 cluster. Three out of four measures had a higher loading on the speaking factor compared to the loading of the same measures on the general factor in the three to five cluster which is of concern in this model. The general factor does not seem to explain as much covariance among the speaking measures for this age group. All the other measures had higher loadings on the general factor compared to the loadings on the domain specific factors. In the 9 to 12 cluster, all of the factor loadings on the general factor were consistently higher than the factor loadings were on the domain specific measures, as expected.

Previous studies have yielded different results about the factor structure of ELP tests; however very few studies have been conducted on the dimensionality of language proficiency tests used in the K-12 setting. Most of the studies have been done on tests used for admission purposes for international students in a college setting. Even though the purpose of the test and the population that takes the test is slightly different, it can be argued that the construct being measured is the same--ELP.

Few studies on the dimensionality of ELP tests have tested the bifactor model as a plausible model. Sawaki et al. (2009) rejected the bifactor model as a plausible model based on the factor structure of the Test of English as a Foreign Language (TOEFL). The finding from this study that the bifactor model can best explain the data is contradictory to the TOEFL study which reports the bifactor model as an implausible one due to non-significant and low loadings on the general factor. Contrary to the findings by Sawaki et al. (2009), the bifactor loadings for this study were higher on the general language factor which shows that there is a factor that explains English language proficiency as measured by the ELDA over and above the four-factors of reading, writing, listening, and speaking.

It is of importance to note that the bifactor model in Sawaki et al.'s (2009) TOEFL study was specified differently than in this study. In the current study, the correlations between the factors were set to a zero in the bifactor model, whereas in the TOEFL study the factors were not

constrained to zero. The other difference between the two studies is that the analysis for the TOEFL study was done at the item level but in this study the measures were item parcels aggregated based on the standard being measured.

Second-order Factor Model

The results from the second-order model suggested adequate fit for the data for both the three to five grade cluster and the 9 to 12 grade cluster. The factor loadings for the general second-order factor as well the four lower factors reading, writing, listening, and speaking were all high and significant. The results were consistent for both grade clusters. The standardized factor loadings on the general language proficiency factor were very high for reading (0.91 for both grade clusters) and writing (0.99 for grades three to five and 0.97 for grades 3 to 12) followed by listening (0.87 for the three to five and 0.92 for the 9 to 12 grade clusters). Similar to the bifactor model results, speaking had the lowest loading (0.67 for three to five and 0.75 for 9 to 12) on the general language factor. This indicates that the second-order English language proficiency factor has the four hypothesized underlying dimensions: reading, writing, listening, and speaking.

The loadings on the lower order factors were consistently higher in the 9 to 12 cluster for all measures. The factor loadings on the lower order factors (reading, writing, listening and speaking) were high indicating the four dimensions are distinct constructs being measured. This was the

same for both the three to five and the 9 to 12 grade clusters. There were some inconsistencies in the trends for the factor loadings between the two grade clusters. The highest loadings among the lower order factors were in the speaking factor for the measure 'expand' was 0.93 and for 'tell' was 0.92 in the 9 to 12 cluster. In the three to five grade cluster, the highest loading was on the reading measure 'identifying supplemental ideas' at 0.88 and the second highest was on the speaking measure 'expand' at 0.86.

The results from this study about the second-order factor structure is consistent with the previous studies which report that language has multiple components and that it has both a second-order factor and domain specific factors (Bachman & Palmer, 1982; Sawaki et al., 2009; Shin, 2005). The second-order model in this study fit the data well, which indicates that there is a common underlying dimension or factor across the four domains of reading, writing, listening and speaking. Bachman and Palmer (1982) concluded that there was a general order factor and two first order factors, sociolinguistic competence and grammatical/pragmatic competence. Shin (2005) reported a second-order factor and three domain specific factors of listening, written expression, and speaking as the factor structure for the older version of the TOEFL test combined with the Test of Spoken English. The reading comprehension measures in this model were allowed to load on the written expression factor. This study was consistent with the finding that there is a second-order factor, but

differs in the conclusions about what the first order factors are in all of the other studies except the TOEFL study which concluded there were four first order factors: reading, writing, listening, and speaking.

Correlated Four-factor Model

The correlated four-factor model was a good fit for the ELDA data which tested the hypothesis that there are four distinct dimensions measured. The factor loadings for each of the domains were statistically significant and almost identical to the factor loadings from the lower order factors of the second-order model. The factor loadings for each of the factors-- reading, writing, listening, and speaking-- were high as indicated above, suggesting that the four dimensions are measured by the test even though they are highly correlated.

The results were consistent with the Sawaki et al. (2009) TOEFL study, which reported the correlated four-factor model to adequately represent the data. The TOEFL study reported the fit of the correlated four-factor model to be comparable to the second-order model, but since the second-order model was parsimonious, it was chosen as the best model to represent the TOEFL test. It is of interest, however, that in the TOEFL study, the loadings on the bifactor model for each of the domains were identical to the four-factor model, but in this study the factor loadings of the second-order model and the four-factor model are nearly identical for each of the domains.

The factor structure of the ELDA seems most comparable to the factor structure of the Internet-based TOEFL test where the researchers concluded that the second-order factor model and the four domain specific factors of reading, writing, speaking, and listening best explains the data (Sawaki et al., 2009). The findings from the Internet-based TOEFL study (2009) were consistent with the finding from this study that the four-factor correlated model fit the data well.

The results from this study contradict the findings from some of the previously done studies in this area, and this speaks to the complexity of the operational definition of the construct of language proficiency. The study done by Stricker et al. (2005) analyzing LanguEdge--a language proficiency test which is very similar to the TOEFL test with the four sections of reading, writing, listening, and speaking--showed different results than this study and the TOEFL study. Stricker et al. (2005) found two-factors, a speaking factor and a factor that combines listening, reading, and writing. This is very different, because they did not find four distinct constructs being measured on this instrument. As mentioned in the discussion, the speaking factor is the least correlated with the other factors in the different models in this study as well, and a two-factor model was tested which is discussed below.

Two-factor Model

This study also examined a two-factor solution where the distinction was made between the skills that are acquired, listening and speaking,

and the skills that are learned, reading and writing. The model did not fit the data well, and hence was not considered a plausible model. However, there is indication in the data that speaking is a different construct compared to the other constructs measured by the test. Speaking is the least correlated with the other factors in all models which suggest that there may be other models that fit the data better than the models tested in this study. This study did not find favorable results with the two-factor model though factor loadings were moderately high for factors, language (speaking and listening), and literacy (reading and writing). The results were consistent in both grade clusters and similar to the results in other models. The factor loadings in the 9 to 12 grade cluster were higher than three to five for both factors. The correlation between the factors was high at 0.79 for the three to five grade cluster and at 0.79 in the 9 to 12 cluster. The factor loadings were significant for all measures in both factors. Examining the factor loadings, it is unclear why the model did not fit the data.

Single Factor Model

The results from the single factor model do not indicate that language proficiency is a unidimensional construct. The fit indices indicated poor fit for the data, but the TOEFL study (Sawaki et al., 2009), reported the fit indices to be acceptable, but the model was rejected because the chi-square difference test indicated a much better fit for the four-factor correlated model in their study. The single factor model had

the lowest CFI and RMSEA (greater than 0.1 for both the three to five cluster and 9 to 12 clusters) compared to all the models tested in this study. The factor loadings were all significant and on the moderate side (0.54-0.78 for three to five and 0.50- 0.84 for 9 to 12 cluster) compared to the other models. This is consistent with the findings from other studies where ELP is considered as multidimensional with a second-order factor and domain specific factors.

Conclusions

The factor structures reported from the different ELP tests seems to have similar yet different factor structures. This could be explained by differences between these studies and the measures used for the analyses. In this study, item parcels were used, whereas the TOEFL study used individual items. The TOEFL test was also an Internet-based test for students entering college, and the ages and the language backgrounds of these students could be very different from the sample in this study. The language abilities of these students may be also very different based on the students in this sample. Stricker et al. (2005) used item parcels similar to this study, but reported a two-factor solution which suggests that there are only two constructs (a speaking factor and a factor that combines listening, reading, and writing) being measured by the test even though it has four different subtests, reading, listening, reading, and writing.

Even though the tests measure language proficiency, ELDA is a test that measures English language proficiency based on the standards,

whereas TOEFL is designed to measure whether students can successfully communicate in college. Even though there is a significant overlap in the definition of English language proficiency for these two tests, the test may be measuring slightly different constructs because the purpose of the test is different. However, this is an important issue and should be carefully considered when making decisions about the scoring and interpretation of test scores from an ELP test.

The consistency in the factor structures of the two grade clusters examined provided some support for the argument that the test measures the same constructs at both levels, although further tests of measurement invariance between these groups are required to generalize this finding. The test is designed to measure growth, and it is important that the constructs measured are the same. This study suggests that for both grade clusters tested the factor structure is consistent and that it tests the four domains of the reading, writing, speaking, and listening separately and that ELP is not a unidimensional construct. Even though this study suggests that the bifactor model is the best fit for the data, this has to be tested using other ELP tests to ensure that this is replicable.

Conjunctive versus compensatory scoring. Dimensionality of the test is important to this discussion of how proficiency levels are determined for each domain, as well as the classification levels in the overall category. The practice of reporting an overall proficiency score suggests that ELP can be considered as a unidimensional construct even

though this study does not support that conclusion. Central to this discussion is also how the scores are combined to create an overall proficiency level. For the ELDA, the proficiency levels for reading, writing, listening and speaking were determined by committee recommendations and the overall proficiency levels were determined by combining the different proficiency levels.

The four domains for the ELDA are highly correlated, and this suggests that combining the different scales to form a single construct is warranted. ELDA combines the score from each of the domains by combining the levels based on the recommendations from the technical advisory committee. The comprehension score is weighted more heavily towards reading. If the student gets a 5 (fully English proficient) on listening and a 1 (pre-functional) on reading, the comprehension score is a 2 (beginner). In the production levels, writing is weighted heavily. The composite is a combination of comprehension and production levels and hence, reading and writing are given more weight than listening and speaking.

The rules used to weight the scores are important to this discussion of dimensionality. Abedi (2007) states that the researchers should ask,

> Should the four domains be considered as four separate subscales/dimensions or should they be considered as a single latent trait that encompasses all four domains?" There are different models and different views on this choice. The number of

> constructs being measured seriously affects reporting and interpretation of scores. If the four domains are measuring a single construct (i.e., the overall English language proficiency latent variable) then scores from the four domains can be combined and a single score can be used for reporting Annual measurable achievement objective and for classification purposes. On the other hand, if each domain has a unique contribution to the ELP construct, how can a total score be obtained and interpreted? (p.125)

This study suggests that there are four distinct domains which contribute to the factor structure of the ELDA and a combination of scores where reading and writing are weighted more heavily is not warranted. This study suggests that all the four domains should be weighted equally in reporting a composite score. The results of the bifactor model indicate that there are specific contributions from each of the constructs of reading, writing, listening and speaking over and above the general factor, and this should be considered in scoring decisions. ELDA uses a compensatory weighted model which gives more weight to the reading and writing tests which is of concern that all domains are not considered equal.

The correlation between the reading and writing factor is very high which indicates that they can be combined when making scoring decisions, but speaking in not correlated that highly with the other measures which means that it is a different ability compared to the other

constructs measured using ELDA. However, it is not one of the constructs weighted heavily in the identification of the composite level. It is also of interest that reading is combined with listening to get a level and writing is combined with speaking. The results from this study indicate that reading and writing are very highly correlated and in compensatory scoring the two domains could be combined to get an overall score. Reading and writing are correlated moderately with speaking which warrants compensatory scoring, but speaking should be treated as a separate construct which is consistent with other studies that has reported a distinct speaking component. The two-factor model that was tested in this study did not yield a good fit but the correlations between factors indicated that this factor structure should be investigated more with different ELP tests to ensure the accuracy of that solution.

The dimensionality of the test and the exact structure of the constructs being measured have an impact on the reporting, interpretation and the decisions made about the use of the scores. ELP tests are high stakes tests because many decisions about program placement and the type of services received by ELL students are determined by the score from the ELP test exclusively in some states or in conjunction with other measures in most states. Different states (Arkansas, Iowa, and Louisiana) use the ELDA to identify ELL students as the only criterion (Wolf et al., 2008). The technical report on the standard setting process does not

provide a rationale on the score combining rules established by the expert committee (Bunch, 2006).

The inaccuracy of the proficiency levels used for classifying students as ELL or non-ELLs may lead to inadequate and ineffective instruction of ELL students in program, and this may negatively impact their schooling. There have been validity concerns about the ELL identification and reclassification practices and that different ELP tests have used different criterion to be identified/reclassified from the program (Abedi, 2007). The proficiency level reported by the test has high stakes consequences for ELL students, and hence a strong rationale should be established for determining overall proficiency levels. This high stakes decision about overall proficiency level should also incorporate the second language acquisition theory about the level of second language required for students to effectively participate and keep up with language demands of school.

In the standard setting process the content area experts should be presented with the data about the dimensionality of the constructs being measured. The results from the study recommend conjunctive scoring where the student should be proficient in all domains to achieve overall proficiency. If the content area experts choose composite scoring a rationale should be given based on the dimensionality of the test so that practitioners can make informed decisions about the proficiency level classifications from the ELP test. The results from this study are limited to

the ELDA and cannot be generalized to other ELP tests because there was no consistent definition for the construct of ‘academic English’ that all ELP tests intends to measure.

Limitations. There are some major limitations to this study. The data set did not provide demographic information about the students and the sample may not be representative of the ELL students in K-12 settings in the United States. The factor structure may differ based on the different language backgrounds and experiences represented in the sample. This information was not available in the data set, and hence no conclusions can be made about that. The study had very few students who were classified as proficient based on the overall score which is a major limitation because studies in the past have reported changes in the interpretability of factor structure based on ability levels (Davidson, 1995; Swinton & Powers, 1980). The students were clustered in the three proficiency levels in the middle in this sample which makes the sample homogenous in ability levels.

The results shed light on the factor structure of ELP as measured by ELDA, which provides validity evidence. But future efforts should validate the use of the score and the proficiency levels reported by ELDA by comparing the performance of the students classified at different levels against other reliable and valid measures. It would be of great value if the study is replicated with other ELP tests that are used in schools today. This would inform policy makers and test developers to make better

decisions about the use of ELP scores. This would also provide valuable insight into the validity of proficiency levels reported by the ELP tests and whether these tests all measure the same construct as mandated by NCLB. Further research has to be conducted to find how valid the proficiency levels reported by the ELDA and other ELP tests are useful for reclassification of ELL students as Fully English proficient by comparing the performance of students exited from the program to the non-ELL students.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*, 4-14.

Abedi, J. (2007). *English language proficiency assessment in the nation: Current status and future practice.* University of California, Davis California.

Abedi, J. (2008). Classification system for English Language Learners: Issues and Recommendations. *Educational Measurement: Issues and Practice, 27*(3), 17-31.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.

American Institutes for Research. (2005). *English Language development Technical Report (ELDA): 2005 Operational and field test administrations.* Council of Chief State School Officers.

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

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language: A Cambridge- TOEFL comparability study.* Cambridge, England: Cambridge University Press.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document.* Los Angeles, CA: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Bauman, J., Boals, T., Cranley, E., Gottlieb, M., & Kenyon, D. (2007). Assessing comprehension and communication in English state to state for English language learners (ACCESS for ELLs). In J. Abedi (Ed.), *English language proficiency in the nation: Current status and future practice* (pp. 81-92). Davis, CA: University of California.

Bialystok , E., & Hakuta, K. (1994). *In other words: The science and psychology of second language acquisition.* New York, NY: Basic Books.

Bialystok , E., & Martin, M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science, 7*(3), 325-339.

Bridgeman, B., & Harvey, A. (1998). Issues in developing and administering a test of English language proficiency. *Presented at the National Council on Measurement in Education.* San Diego: ERIC document.

Bunch, M. B. (2006). *Final Report on ELDA standard setting.* Measurement Incorporated.

Bunch, M. B. (2011). Testing English language learners under No Child Left Behind. *Language Testing, 28*(3), 323-341.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Chomsky, N. (1959). Review of verbal learning. *Language, 35*, 26-58.

Chomsky, N. (1975). *Reflections on Language.* New York, NY: Pantheon Books.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin and use.* New York, NY: Praeger.

Chomsky, N. (1995). *The minimalist program.* Cambridge, MA: M.I. T Press.

Crystal, D. (1986). The prescriptive tradition. In D. Crystal (Ed.), *The Cambridge encylopedia of Language.* Cambridge, England: Cambridge University Press.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 221-251.

Cummins, J. (1981). *Bilingual and minority-language children.* Toronto, Canada: OISE Press.

Cummins, J. (1984). Language proficiency and academic achievement revisted: A response. In C. Rivera (Ed.), *Language proficiency and academic achievement* (pp. 71-76). Clevedon, UK: Multilingual Matters.

Cummins, J. (2000). Putting language proficiency in its place: Responding to critiques of the conversational/academic language distinction. In J. Cenoz, & U. Jessner (Eds.), *English in Europe: The acquisition of a third language.* Clevedon,UK: Multilingual Matters.

Davidson, F. (1995). Language test unidimensional model fit at multiple ability levels. (ED 390267). Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED390267&ERICExtSearch_SearchType_0=no&accno=ED390267

Del Vecchio, A., & Guerrero, M. (1995). *Handbook of English language proficiency tests.* Albuquerque, NM: New Mexico Highlands University.

Duncan, S. E., & DeAvila, E. A. (1979). Bilingualism and cognition: Some recent findings. *NABE Journal, 4*(1), 15-50.

Edelsky, C., Hudelson, S., Flores, B., Barkin, F., Altwerger, B., & Jilbert, K. (1983). Semilingualism and language deficit. *Applied linguistics, 4*, 1-22.

Fabrigar, L. R., MacCallum, R. C., Strahan, E. J., & Wegener, D. T. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quarterly, 16*(1), 43 to 59.

Fast, M., Ferrara, S., & Conrad, D. (2004). *Current efforts in developing English language proficiency measures as required by NCLB: Description of an 18-state collaboration.* Washington, DC: American Institute for Research.

Forte, E. (2007). *How states are defining, measuring, and evaluating proficiency among English language learners.* Washington DC: Council of Chief State School Officers.

Gandara, P., & Hopkins, M. (2010). The changing linguistic landscape of the United States. In P. Gandara, & M. Hopkins (Eds.), *Forbidden language: English learners and restrictive language policies* (pp. 7-19). New York, NY: Teachers College Press, Columbia University.

Garcia, E. E., Lawton, K., & Diniz de Figueiredo, E. H. (2010). *Assessment of young English language learners in Arizona: Questioning the validity of the state measure of English Language Proficiency.* Los Angeles, CA: The Civil Rights Project.

Gorsuch, R. L. (1983). *Factor analysis.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. (2004, March). Construct-irrelevant variance in high stakes testing. *Educational measurement: Issues and practice, 23*(n1), 17-27.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002, November). The faculty of language: What is it, who has it, and how did it evolve. *Science's Compass, 298*, 1569-1579.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling, 6*, 1-55.

Kessler, C., & Quinn, M. E. (1982). Cognitive development in bilingual environments. In B. Hartford, A. Valdman, & C. R. Foster (Eds.), *Issues in international bilingual education: The role of the vernacular.* New York, NY: The Plenum Press.

Kim, J.-O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues.* Newbury Park, CA: Sage Publications.

Kunnan, A. (1995). *Test taker characteristics and test performance: A structural equation modeling approach.* Cambridge, UK: Cambridge University Press.

Lara, J., Ferrara, S., Calliope, M., Sewell, D., Winter, P., & Kopriva, R. (2007). The English language development assessment (ELDA). In J. Abedi (Ed.), *English language proficiency assessment inthe nation: Current status and future practice* (pp. 47-60). Davis, CA: University of California.

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English language learners.* Santa Barbara, CA: Linguistic Minority Research Institute.

MacSwan, J. (2000). The threshold hypothesis, semilingualism, and other contributions to deficit view of linguistic minorities. *Hispanic journal of behavioral sciences, 22*, 3-45.

MacSwan, J., & Rolstad, K. (2003). Linguistic diversity, schooling, and social class: Rethinking our conception of language proficiency in language minority education. In C. B. Paulston, & R. Tucker (Eds.), *Sociolinguistics: The essential readings* (pp. 329-341). Malden, MA: Blackwell.

MacSwan, J., & Rolstad, K. (2010). The role of language in theories of academic failure for linguistic minorities. In J. E. Petrovic (Ed.), *International perspectives on bilingual education: Policy, practice, and controversy* (pp. 173-194). Charlotte, NC: Information Age Publishing.

Mahoney, K., & MacSwan, J. (2005). Reexamining identification and reclassification of English language learners: A critical discussion of select state practices. *Bilingual research journal, 29*(1), 31-42.

Mahoney, K., Haladyna, T., & MacSwan, J. (2009). The need for multiple measures in reclassification decisions: A validity study of the Stanford English Language Proficiency Test (SELF). In J.S. Lee, T.G. Wiley, & R. Rumberger (eds.), *The education of language minority immigants in the USA* . Bristol, UK: Multilingual Matters.

Martin-Jones, M., & Romaine, S. (1986). Semilingualism: A half baked theory of communicative competence. *Applied linguistics, 7*(1), 27-38.

Mulaik, S. A. (1972). *Foundations of factor analysis.* New York, NY: McGraw Hill.

National Clearinghouse for English Language Acquisition. (2011). *The growing numbers of English language learner students (from 1998-99 to 2008-09).* Retrieved from http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0809.pdf

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115.

Pinker, S. (1994). *The language instinct: How the mind creates languages.* New York, NY: William Morrow.

Reymont, R., & Joreskog, K. G. (1993). *Applied factor analysis in the natural sciences.* New York, NY: Cambridge University Press.

Romaine, S. (1995). *Bilingualism* (2nd ed.). Cambridge: MA: Blackwell Publishers.

Salehi, M. (2011, June). On the factor structure of a reading comprehension test. *English language teaching, 4*(2), 242-249.

Salehi, M., & Rezaee, A. A. (2009). On the factor structure of the grammar section of University of Tehran language proficiency test. *Indian journal of applied linguistics, 35*(2), 169-187.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet based test. *Language testing, 26*(1), 5-30.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*(1), 31-57.

Slobin, D., & Bowerman, M. (1985). Crosslinguistic evidence for the language making capacity: What shapes children's grammar? Hillsdale, NJ: Erlbaum.

Skinner, B. F. (1957). *Verbal behavior.* New York, NY: Appleton.

Standards for educational and psychological testing. (1999). Washington DC: American Educational Research Association.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge Test across language groups.* Princeton: NJ: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the Test of English as a Foreign language for several language groups.* Princeton, NJ: Educational Testing Service.

Tate, R. (2002). Test dimensionality. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity,*

technical adequacy and implementation. (pp. 181-211). Mahwah, NJ: Lawrence Erlbaum Associates.

Title III : English Language Acquisition, Language Enhancement, and Academic Achievement Act. 20 U.S.C. 6801 et seq. (2004).

Valdes, G., & Figueroa, R. A. (1994). *Bilingual and testing: A special case of bias.* Norwood, NJ: Ablex.

Wiley, T. (1996). *Literacy and language diversity in the United States.* McHenry, IL: Center for Applied Linguistics and Delta System.

Wolf, M. K., Kao, J. C., Herman, J., Bachman, L. F., Bailey, A. L., Bachman, P. L., et al. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses.* Los Angeles, CA: University of California.

Zehler, A. M., Hopstock, P. J., Fleischmann, H. L., & Grenuik, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Special Issues Analysis Center.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Marilyn Thompson EDB
From:	Mark Roosa, Chair Soc Beh IRB
Date:	10/07/2011
Committee Action:	**Exemption Granted**
IRB Action Date:	10/07/2011
IRB Protocol #:	1109006898
Study Title:	Language Proficiency testing

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(4) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.