Fall March 2016

Examining the Reliability and Validity of ADEPT and CELDT: Comparing Two Assessments of Oral Language Proficiency for English Language Learners

Gina Chavez
Loyola Marymount University

Fall October 2013

Examining the Reliability and Validity of ADEPT and CELDT: Comparing Two Assessments of Oral Language Proficiency for English Language Learners

Gina Chavez
Loyola Marymount University, ginachavez@mac.com

LOYOLA MARYMOUNT UNIVERSITY

Examining the Reliability and Validity of ADEPT and CELDT:

Comparing Two Assessments of Oral Language Proficiency for English Language Learners

by

Gina Chavez

A dissertation presented to the Faculty of the School of Education,

Loyola Marymount University,

In partial satisfaction of the requirements for the degree

Doctor of Education
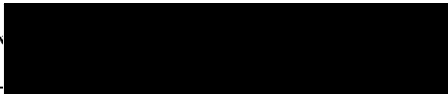
2013
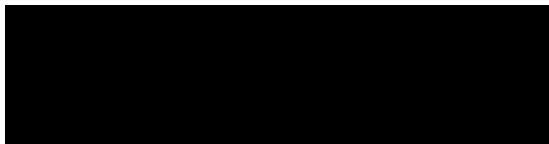
Loyola Marymount University
School of Education
Los Angeles, CA 90045
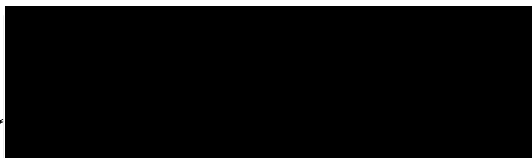
This dissertation written by Gina Chavez, under the direction of the Dissertation Committee, is approved and accepted by all committee members, in partial fulfillment of requirements for the degree of Doctor of Education.

9-1-2013
_____
Date

Dissertation Committee

_____
Magaly Lavadenz, Ph.D., Committee Chair

_____
Marta Sanchez, Ph.D., Committee Member

_____
Karen Huchting, Ph.D., Committee Member

ACKNOWLEDGEMENTS

DEDICATION

To my parents, Rudy and Carmela Chavez, for bringing me into this existence even though the odds were against it. For all the sacrifices you made to make my life experience better than yours and for always telling me "You can do it, mija!"

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Few classroom measures of English language proficiency have been evaluated for reliability and validity. Researchers have examined the concurrent and predictive validity of an oral language test, titled A Developmental English Language Proficiency Test (ADEPT), and the relationship to the California English Language Development Test (CELDT) in the receptive/listening and expressive/speaking domains. Four years of retroactive data representing 392 student records were obtained from a local urban school district in Los Angeles County with a significant proportion of English language learners. After preparing the data file for analysis, data was analyzed using the Statistical Package for the Social Sciences (SPSS) system. Cronbach's alpha was used to analyze the internal consistency of ADEPT. Pearson r analysis was performed to examine concurrent validity and predictive validity. Findings indicated moderate to high correlation coefficients of internal consistency in the first three levels of ADEPT. Concurrent validity results varied depending on the school year. In the most recent school year, 2012–2013, positive moderate to strong correlations were found. This relationship was weaker in each previous year. Overall, correlations increased and remained positive as sample size increased but predictive validity was weak for all three sets of comparative years. These findings support the use of ADEPT as a multiple measure, as a monitoring tool and to inform instruction.

CHAPTER 1

INTRODUCTION

Background

In last decade, the United States has seen unprecedented focus on the academic

achievement of English language learners (ELLs) due to the 2002 reauthorization of the

Elementary and Secondary Education Act (ESEA) also known as the No Child Left Behind Act

NCLB, 2002,. ELLs—also referred to as English learners (ELs) in the literature—are students

still learning English but in classrooms where instruction is delivered primarily in English. This

federal law holds local educational agencies accountable not only for the language development

of ELL students but also for their achievement in reading/language arts and mathematics. In the

past, ELLs were often excluded from accountability systems because they entered school without

the requisite English language skills to benefit from the mainstream curriculum (Bailey, 2007).

In addition, educators and school systems have often overlooked oral language, which includes

the language domains of listening and speaking, and its assessment for educational purposes.

Bailey (2010), in a review of publicly available documents published by the 50 United States and

the District of Columbia, found that 41 of their education agencies do not include an oral-

language component in their state-wide tests of English language arts (ELA), despite the fact that

48 of the agencies include oral-language skills in their mandated language-arts-content standards.

These mandated ELA content standards are the year-end goals for all students, which ELLs must

achieve to proficiency in addition to learning English proficiently.

Under the Title III law of NCLB, state and district progress of ELLs is measured against

state-established annual measurable achievement objectives (AMAOs), which every district is

responsible for meeting annually. The objectives are: (a) AMAO 1: Annual increases in the number or percentage of students showing progress in learning English; (b) AMAO 2: Annual increases in the number or percentage of students attaining English proficiency; and (c) AMAO 3: Making adequate yearly progress for limited English proficient children as described in Title I, Section 1111(b)(2)(B), of ESEA. The legislation states that data must be used to improve educational outcomes for English learners and ultimately "to stimulate activities that better support ELs and increase English proficiency and content knowledge" (Tanenbaum & Anderson, 2010, p. 3).

In 1998, the placement of ELLs was dramatically impacted by the passage of Proposition 227, designed to eliminate instruction in primary language. The passing of this proposition proved to be a defining moment for ELLs because prior to this initiative, ELLs were tested in their primary language to determine their academic achievement; they were also tested for proficiency in English. At present (2013), ELLs take the California state-wide test for academic achievement in English even though they are still learning the language.

Assembly Bill 748, approved by the State Board of Education (SBE) in 1999, mandated the creation of the English Language Development (ELD) Standards, triggered by Proposition 227 (California Department of Education, 2002). The function of the ELD standards was to mainstream all ELLs into the regular language arts curriculum. The ELD standards supplement the English–language arts content standards to ensure that ELLs develop proficiency in both the English language and the English–language arts content standards (California Department of Education, 2002).

The same legislation required the SBE to identify or develop a test to assess the language

proficiency levels for English Learners. As a result, the California English Language

Development Test (CELDT) was created, and field-testing took place in Fall 2000 with a

volunteer population of schools (California Department of Education, 2011). The first edition of

CELDT, used state-wide, was in 2001–2002 and continues to be used for the Title III

requirement to monitor ELL progress in the acquisition of English. This assessment is

administered once a year, providing summative information, which is how well students are

progressing in reaching standards (Gándara & Rumberger, 2007). Summative information

obtained at the end of the school year is primarily used for evaluating student performance and

program efficacy (Bailey, 2010). This type of assessment is not designed to inform instruction

during the school year.

　　This accountability system has created a dual challenge for ELLs, as they are being

tested to ascertain their level of proficiency in English and their achievement in the content

standards. Because they are given annually, the tests provide a single opportunity for ELLs to

show growth in English proficiency and the sole opportunity to demonstrate achievement toward

the content standards.

　　Teachers in California have identified a dilemma with the use of CELDT results because

they do not provide information to help improve instruction for ELLs. A survey conducted by

Gándara, Maxwell-Jolly, and Driscoll, 2005, with nearly 5,300 California educators responding,

found that "The CELDT . . . does not provide them a great deal of useful information of a

diagnostic nature" (p. 9). This finding is troubling because there are no other "mandated"

language tests for ELLs in California, and the CELDT was not designed to be instructionally

diagnostic. Indeed, it was intended to identify ELLs, to determine level of English proficiency,

and to assess annual progress toward English proficiency. Although, districts are required to use multiple criteria for the purposes of reclassification, only the CELDT is mandated to demonstrate annual progress (Grissom, 2004). In addition, "Although CELDT is a standards-based measure of English proficiency, it is only one assessment, and its current test window does not provide timely information for decisions on instruction or redesignation" (Parrish et al., 2002). Districts that use only CELDT data for decision-making continue to receive limited information about progress in English proficiency, thereby reducing the chances that instruction is modified for ELLs.

The purpose of this study was to measure the psychometric properties of an oral language assessment instrument that can be used by teachers during the school year and which provides instructional information regarding a student's English proficiency level. A Developmental English Language Proficiency Test (ADEPT) is an oral language test of listening and speaking skills. It can be used with all ELLs two to three times per year. According to the assessment manual, "ADEPT assess a student's ability to understand and generate utterances using a scope and sequence of language forms, or structures, across the five levels of English proficiency (California Reading and Literature Project, 2006, p. 2). Moreover, "Test results demonstrate what students can produce naturally based on the knowledge of English they have internalized. It reveals a student's command of key grammatical structures within each of the levels of proficiency" (p. 4).

The Alisal Union School District, located in Salinas, California, created the first edition of ADEPT. This district used it for over 20 years before the California Reading and Literature Project (CRLP) was granted permission to develop an adaptation in 2000. The CRLP conducted

4

field-testing with thousands of teachers and students from 2000–2003. In 2004, the University of California at Santa Cruz conducted a reliability and validity study of ADEPT and found it to be a reliable and valid instrument that aligned with the CELDT. As a result of this study, the CRLP published the current version in 2006.

ADEPT assesses the skills of listening and speaking by focusing on the use of specific grammatical forms (California Reading and Literature Project, 2006). ADEPT is administered to students individually according to their proficiency level. Students are given general questions or prompts related to a picture they are viewing. The tester uses carefully worded prompts to elicit student responses. Student responses reveal "valuable information and insight into a student's command of each grammatical structure and the related general utility vocabulary upon which the use of that structure depends"(California Reading and Literature Project, 2006, p. 2). ADEPT assesses receptive (listening) and expressive (speaking) English language proficiency at the first three levels of proficiency. The fourth level assesses only expressive proficiency.

If the results from 2004 are replicated, districts could use ADEPT with confidence as an additional measure for monitoring progress toward English language proficiency. With the ADEPT results, teachers can increase their knowledge of grammatical structures that are not only challenging for ELLs but also prevent them from acquiring English at a proficient level. Further, teachers can adjust their instruction accordingly and develop improved strategies to meet the needs of ELLs.

<div align="center">

Statement of the Problem

</div>

The CELDT is a summative assessment administered to identify students who are limited-English-proficient to determine their levels and to assess their progress in acquiring

English language proficiency (California Department of Education, 2011).. By definition, summative assessments do not provide information to support instruction during the year. Summative assessment "provides useful information . . . of students' achievement or progress at the end of a course of study" (Bachman & Palmer, 1996, p. 98). Consequently, no mandated tests in California monitor ELL progress toward English proficiency during the school year.

The test investigated for this study is called A Developmental English Language Proficiency Test (ADEPT) and has been used by a Northern California school district for over 30 years. In 2004, the UC Santa Cruz Educational Partnership Center conducted studies to establish the reliability and validity of ADEPT as well as its alignment to the CELDT. The study found a correlation of .76, indicating good concurrent validity between these two tests and suggesting that ADEPT captures a good portion of the language proficiency skills assessed by CELDT in the listening and speaking domains.

In 2006–2007 the CELDT underwent revisions to ensure the test was federally compliant with NCLB legislation. For this reason, the reliability and validity of ADEPT and concurrent validity to the CELDT must be replicated.

Purpose of the Study

The purpose of this study was to determine the reliability and validity of ADEPT, adapted by the California Reading and Literature Project (CRLP), an oral language test of listening and speaking. This test can be used with all ELLs two to three times per year to monitor ELL progress toward English proficiency during the school year. Currently, no valid form of assessment provides practical guidance for teachers to inform instruction for ELLs. Districts/schools that include ADEPT and CELDT in their assessment systems are going beyond

what is required by California law. Therefore, it is critical to document the reliability and validity of ADEPT and the correlation to CELDT.

Significance of the Study

Due to the limited information provided by the CELDT, and the fact that it is an annual assessment, teachers need additional data to understand how well ELLs are progressing toward English proficiency during the school year. Because CELDT is administered at the end of the school year, the current grade-level teacher cannot influence the performance of their ELLs on the annual test. ADEPT can provide progress information during the school year when a teacher has the best opportunity to influence instruction toward English proficiency.

A compounding factor is that teachers may lack the pedagogical understanding of English language development and the practical application to meet the needs of ELLs even if they have required certification in this area. Although a teacher may be qualified to teach a content area such as English language arts or mathematics, "the same teacher may not possess the skills, experience, or pedagogical practices to teach students whose native language is not English" (Cadiero-Kaplan & Rodriguez, 2008, p. 376). This finding illustrates a potential gap in what teachers can provide instructionally and what ELLs need to progress. Therefore, teachers need specific information about English language proficiency skills and students' language abilities. Thus, providing teachers tools such as ADEPT to support their understanding of English language proficiency so they can adjust their instruction accordingly is imperative.

Generally, research in the area of classroom assessment is not lacking, but a paucity of research exists when it comes to English language proficiency, specifically the role of English language development standards as the basis for classroom assessment (Llosa, 2011). Although

the CELDT is based on the 1999 English language development standards, student performance on a general language test of this type does not guarantee attainment of the necessary language skills for academic tasks. Thus an important assessment gap exists between the type of English an ELL may know and be able to use and the language critical to school success (Bailey, 2007).

Theoretical Framework

The Equal Protection Clause of the 14th Amendment to the US Constitution protects individuals' rights to an equal educational opportunity. This opportunity cannot be achieved merely by providing all students with the same facilities, textbooks, teachers, and curriculum, because students who do not understand English are effectively foreclosed from any meaningful education. Therefore, it is the responsibility of every school district to take affirmative steps to overcome educational barriers faced by non–English speaking students. The Supreme Court made this ruling in 1974 in the landmark case of Lau v. Nichols (Hakuta, 2011). Prior to this case, the first official federal recognition of the needs of students with limited English speaking ability (LESA) was the Title 1 of the Elementary and Secondary Education Act (ESEA) of 1965. Title 1 sought to "ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments" (http://www2.ed.gov/policy /elsec/leg/esea02/pg1.html).

These first two legal milestones eventually led to a third federal ruling, in the case of Castaneda v. Pickard, 1978, which expanded on Lau v. Nichols (1974) in establishing guidelines to judge compliance with the Equal Education Opportunities Act. According to this ruling, if federal funds were used for English learner programs, the program must meet three criteria: (a)

be based on sound educational theory or principles; (b) effectively implement this theory; and (c) produce results indicating that it is working (Parrish et al., 2006). The Lau v. Nichols (1974) mandate is still in effect and continues to be challenged by state initiatives.

These mandates were integrated into the current NCLB legislation. According to Borkowski and Sneed (2006), the principal benefits of NCLB reside in its recognition that all children can learn and have a right to be taught. Secondly, all stakeholders should be notified of how students are progressing toward the attainment of high academic standards. The educational goals for all students are to reach proficiency on state content standards. English language learners (ELLs) must also reach proficiency in English. To measure if these goals are being met, students are annually tested on state content standards, and ELLs are also tested for English proficiency. Consequently, three theories were examined as part of the theoretical framework for this study: testing theory to examine test creation, use, and purpose; second language acquisition theory to understand how a second language is learned; and sociocultural theory as a theory of learning because a goal of the current accountability system is to measure learning.

Testing Theory

Testing theory provided the basis for understanding test creation including principles for ethical and fair test design. One theory explained testing in terms of how people perform on a test, in contrast to another theory that examined how people respond to test items (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999). These are the two most commonly used theories in educational testing.

The link between testing and academic performance is well documented in the literature. As stated previously, ELLs are tested to determine their academic performance, although they are still learning English, which can put their performance in jeopardy. In California, academic performance is determined by a norm-referenced test known as the California Standards Test (CST). This reality presents additional concerns for ELLs, because norm-referenced and standardized tests continue to be used today for accountability purposes without critical examination of how the results impede equity in our schools. For example, intelligence tests have frequently been used as a basis for segregating and sorting students, specifically those whose cultures and languages differ from the mainstream (Neito, 2004). Therefore, the importance of understanding testing theory cannot be underscored given the challenges that ELLs face in our current education system.

Second Language Acquisition Theory

Second language acquisition (SLA) theories are based on learning theories but specifically attempt to explain how one learns a language in addition to one's first language (Lavadenz, 2011). Theoretical frameworks on first and second language acquisition inform instructional methodologies used with ELLs. However, these vary within school districts, which have different program models for reaching their goals. In fact, school districts fulfill the needs of their ELL populations based on their funding and staffing realities. These program models, instructional approaches, and methods used for second language instruction have come under scrutiny due to the accountability mandated in education. The testing of ELLs for English proficiency is now a federal mandate (NCLB, 2002) driven by English language development standards now aligned with the Common Core standards. Understanding second language

acquisition and learning, which is established through testing, is fundamental to the identification, diagnosis, and placement of ELLs in our educational system.

Sociocultural Theory

Sociocultural theory (SCT) suggests that learning is influenced by social, cultural, and historical factors. Learning takes place within social interactions, and teaching occurs through meaningful interactions between experts and novices (Lavadenz, 2011). In contrast to previous learning theories, such as cognitive and behaviorist theories, SCT provides a broader perspective of learning and highlights the importance of language as a tool for mediation. In other words, we use language to learn.

In summary, testing theory explains the principles of test development and the limitations inherent in using test results. Second language acquisition theory delineates the processes of learning a second language. Sociocultural theory explicates learning with the inclusion of language as a tool for learning. SCT is relevant to the present study because of its emphasis on social mediation and interaction, and communicative oral proficiency testing in general (Brooks, 2009).

In the current educational environment, standardized testing has become the measure of success for schools and students. Therefore, there is a need for coherence among learning theory, content instruction—in this case, English language development—and testing; yet a potential gap was found in this relationship. Assessing the language skills of second language learners is recognized as highly complex and prone to biases resulting from a lack of awareness about cultural practices for language usage (Bailey, 2010). Bailey (2010) also found a dearth of research regarding what language demands all students need to master in order to succeed

academically. Consequently, the current accountability measures are limited, which makes it imperative for educators to have a solid understanding of test theory and test-based decision-making.

For purposes of this study, the oral language test that was analyzed focuses on the listening and speaking domains of the English language. This test, titled A Developmental English Proficiency Test (ADEPT), was designed to inform instruction of the English language across the curriculum. The test's focus is on the use of specific grammatical forms and can be characterized as a "classroom assessment of oral language" (Bailey, 2010).

Research Questions

The purpose of this study was to determine the concurrent validity between ADEPT and CELDT. The overarching research questions that guided this investigation were:

RQ1.   What is the reliability of ADEPT?

RQ2.   What is the overall concurrent validity of ADEPT and CELDT in the listening and speaking subscores?

RQ3.   How well does the student's ADEPT score predict the subsequent CELDT?

Research Design and Methodology

The researcher utilized quantitative methods to estimate the reliability and validity of the ADEPT assessment. Retroactive data were collected from a local urban school district that used both ADEPT as a district-wide monitoring tool and the state-mandated CELDT. Scores from both the CELDT and ADEPT were collected; data were analyzed using the Statistical Package for the Social Sciences (SPSS) system. The first question was addressed by using descriptive statistics to analyze the internal consistency of ADEPT. The statistical analysis that was

employed was Cronbach's alpha. To address the second research question regarding concurrent validity, the Pearson *r* analysis was performed. The final question was also measured using Pearson *r* analysis.

Positionality

The positionality of the researcher was that she was involved in the professional development designed to train the teachers in the administration of ADEPT. The researcher had a professional relationship with the district for the past 8 years, which provided some advantage for gaining entry to the district.

Limitations and Delimitations

A limitation to the study was the testing schedule set by the district. The researcher did not have any input regarding the testing schedule; rather, the district followed the recommendations in the administration manual. Additionally, the district decided which ELL students would be tested, so not all ELLs were tested with ADEPT. However, the district's selection criteria focused on the students most in need. The district mandated that ADEPT be administered to students who scored at the beginning and early intermediate levels of proficiency on the CELDT in grades 1–8. These students had the least proficiency in English, as measured by CELDT, and needed to learn English rapidly in order to be competitive on the end-of-year standardized tests. Because ADEPT is administered twice a year, teachers had more timely information about the progress of ELLs toward English proficiency.

The study's delimitations are that it looked only at elementary and middle schools in one urban district. Although the district had seven elementary schools, three middle schools, and a charter high school, the only data used in the study were the results from the 1st- through 8th-

grade students.

Assumptions

Because the data came from a district's data system, two assumptions had to be considered. First, it is assumed that the teachers who administer ADEPT follow the administration guidelines recommended by the authors of the test. Second, it is assumed that the district data system is properly set up to record and retrieve ADEPT and CELDT data.

Definitions of Key Terms

Many of the following definitions are standard definitions from the state or federal department of education. Other definitions provide consistency and are utilized in this research study.

ADEPT: A Developmental English Proficiency Test. A valid and reliable oral language assessment instrument (aligned with the CELDT) that can be used with students across grade levels, K–8 (California Reading and Literature Project, 2006).

AMAO: Annual measureable achievement objectives. There are two annual measurable achievement objectives (AMAOs) for increasing the percentage of EL students making progress in learning English and attaining English proficiency. The third relates to meeting Adequate Yearly Progress (AYP) for the EL subgroup (*Title III Accountability Report Information Guide*, 2009).

API: Academic Performance Index. The cornerstone of California's *Public Schools Accountability Act of 1999*; measures the academic performance and growth of schools on a variety of academic measures (CDE, n.d.).

AYP: A state-wide accountability system mandated by the No Child Left Behind Act of 2001, which requires each state to ensure that all schools and districts make Adequate Yearly Progress (CDE, n.d.).

BICS: Basic Interpersonal Communication Skills; part of a theory of language proficiency developed by Jim Cummins (1984), which distinguishes BICS from CALP (Cognitive Academic Language Proficiency). BICS is often referred to as "playground English" or "survival English." It is the basic language ability required for face-to-face communication, whereby linguistic interactions are embedded in a situational context (see context-embedded language). This language, which is highly contextualized and often accompanied by gestures, is relatively undemanding cognitively and relies on context to aid understanding. BICS is much more easily and quickly acquired than CALP, but is not sufficient for meeting the cognitive and linguistic demands of an academic classroom (Baker & Jones, 1998; Cummins, 1984; NCELA, 2009)

CALP: Developed by Jim Cummins (1984), Cognitive/Academic Language Proficiency is the language ability required for academic achievement in a context-reduced environment. Examples of context-reduced environments include classroom lectures and textbook reading assignments. CALP is distinguished from Basic Interpersonal Communication Skills (BICS) (Baker, 2006; NCELA, 2009).

CELDT: California English Language Development Test. Students in kindergarten through grade 12 whose home language is not English are required by law to take an English skills test. In California, the test is called the CELDT. This test helps schools identify students

who need to improve their skills in listening, speaking, reading, and writing in English. Schools also give the test each year to students who are still learning English (CDE, n.d.).

Construct: The concept or characteristic that a test is designed to measure (American Educational Research Association et al., 1999).

CSTs: California Standards Test. The CSTs are a major component of the Standardized Testing and Reporting program. The CSTs are developed by California educators and test developers specifically for California. They measure students' progress toward achieving California's state-adopted academic content standards in English–language arts (ELA), mathematics, science, and history–social science, which indicate what students should know and be able to do in each grade and subject tested (CDE, n.d.).

CUP: An acronym for Common Underlying Proficiency; Cummins's theory that two languages work in an integrated manner in one underlying, central thinking system. Skills that are not directly connected to a particular language, such as subtraction, using a computer, or reading, may be transferred from one language to another once the concept is understood because they exist as part of the common proficiency. Skills that are specific to a language (idioms, punctuation) may be kept separate (Baker & Jones, 1998). The opposing theory is Separate Underlying Proficiency (SUP) (Baker, 2006, NCELA, 2009).

EL: English Learner (Formerly Known as Limited-English-Proficient or LEP). English learner students are students for whom there is a report of a primary language other than English on the state-approved Home Language Survey *and* who, on the basis of the state-approved oral language (grades kindergarten through 12) assessment procedures and literacy (grades three through 12 only), have been determined to lack the clearly defined English language skills of listening comprehension, speaking, reading, and writing necessary to succeed in the school's regular instructional programs (CDE, n.d.).

ELD: English-Language development is a specialized program of English language instruction appropriate for the English learner (EL) student's (formerly LEP students) identified level of language proficiency. This program is implemented and designed to promote second language acquisition of listening, speaking, reading, and writing (CDE, n.d.).

ELL: English language learner. Students in the process of acquiring social and/or academic English language skills. In most cases, these students have learned a language other than English for use at home or in their community (Burger, Mauricio, & Ryan, 2007).

ESL: English as a second language (ESL) is an educational approach in which English language learners are instructed in the use of the English language. Their instruction is based on a special curriculum that typically involves little or no use of the native language, focuses on language (as opposed to content), and is usually taught during specific school periods. For the rest of the school day, students may be placed in mainstream classrooms, an immersion program, or a bilingual education program. Every bilingual education program has an ESL component (NCELA, 2009)

ESEA: The Elementary and Secondary Education Act, first enacted in 1965, and reauthorized every 5 years. The ESEA was reauthorized as the No Child Left Behind Act in 2001 (NCELA, 2009).

FEP: Students who are fluent-English-proficient are those whose primary language is other than English and who have met the district criteria for determining proficiency in English (i.e., students who were identified as FEP on initial identification and students redesignated from limited-English-proficient [LEP] or English learner [EL] to FEP)(CDE, n.d.).

Formative assessment: Traditionally, the term has referred to assessments used to support learning (Stiggins & Chappuis, 2006).

IFEP: Initially Fluent English Proficient. Refers to a student who is from a language-minority home and who has been determined to be fluent in English upon entering the school system according to a state-approved language proficiency assessment (California Department of Education, YEAR; NCELA, 2009).

LEP: Limited-English-proficient (LEP) students are students for whom there is a report of a primary language other than English on the state-approved Home Language Survey *and* who, on the basis of the state-approved oral language (kindergarten through grade 12) assessment procedures and literacy (grades 3 through 12 only), have been determined to lack the clearly defined English language skills of listening comprehension, speaking, reading, and writing necessary to succeed in the school's regular instructional programs. This term was replaced with the term *English learner* beginning with the 1998–1999 data collection. (CDE, n.d.).

Linguistic minority: A minority language is a language spoken by a minority of the population of a territory. Such people are termed linguistic minorities or language minorities (http://www.museumstuff.com/learn/topics/linguistic_minority).

NCLB: No Child Left Behind. An act intended to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind (NCLB, 2002).

Reliability: The degree to which test scores for a groups of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker (American Educational Research Association et al., 1999).

SLA: Second language acquisition. Refers to both the process and study of developing the ability to use a language other than the native tongue (http://en.citizendium.org/wiki/Second_language_acquisition).

Subgroup: A subgroup is made up of students who share certain characteristics—for examples, students who are economically disadvantaged, students of color, students with disabilities, and students with limited English proficiency (National Education Association, 2008).

Summative assessment: Typically used to evaluate the effectiveness of instructional programs and services at the end of an academic year or at a predetermined time. The goal of summative assessments is to make a judgment of student competency after an instructional phase is complete (Partnership between the Pinellas School District and the Florida Center for Instructional Technology, n.d.).

Validity: The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test (AERA et al., 1999).

Organization of Dissertation

This study determined the concurrent validity of A Developmental English Proficiency Test (ADEPT), an oral language assessment, and the California English Language Development Test (CELDT). Chapter 1 provided a brief outline, including the background, problem, significance, and theoretical framework underlying the research. Chapter 2 reviews the historical background regarding testing, language testing, language acquisition theories, and English language proficiency assessments in the nation, and then specifically in California. Chapter 3 details the methodology employed and the timeline for data gathering. Chapter 4 presents the data analysis of the research. And, finally, Chapter 5 discusses the findings, the significance of the findings and suggestions for future research.

CHAPTER 2

LITERATURE REVIEW

Introduction

This chapter presents the literature related to the testing of English Language learners (ELLs) for English proficiency, particularly in California. The literature review begins with federal and state legislation that defined the educational system in California for all students, including ELLs. The history of education legislation dramatically shifted in the last decade, changing the educational experience for ELLs; but they are still required to achieve academically and learn English at a proficient level. These proficiencies can only be measured through testing to be discussed in this review of the literature on standardized testing and testing theory, including classical test theory and item response theory, which are most widely used in education. Within testing theory is a review of validity theory because this study was aimed at validating an oral language proficiency test of English, titled A Developmental English Proficiency Test (ADEPT) and determining the concurrent validity to the California English Language Development Test (CELDT). Both ADEPT and CELDT measure English language proficiency, thus a review of second language acquisition (SLA) theory is included. The two outcomes for ELLs in the current education system are to learn English at a proficient, academic level and to achieve proficiency on the content standards. Therefore, the theoretical framework for this work includes sociocultural theory, which addresses how students learn; second language acquisition theory, which addresses how to teach ELLs; and testing theory, which addresses how we should measure progress. An ELL student is positioned at the intersection of these three theories, with the educational goals of becoming proficient academically and in English.

Legislation

Federal Legislation

The Equal Protection Clause of the 14th Amendment to the US Constitution protects individuals' rights to an equal educational opportunity. This clause was challenged in court in 1974 in the landmark case of Lau v. Nichols. The Supreme Court ruled that equality of educational opportunity is not achieved by merely providing all students with the same facilities, textbooks, teachers, and curriculum because students who do not understand English are effectively foreclosed from any meaningful education. The court ordered that school districts take affirmative steps to overcome educational barriers faced by non–English speaking students. As a result of this landmark case, dozens of tests were developed to meet the requirement of assessing English proficiency. Some of the tests were developed by researchers or educators and others were naively constructed and are now out of print (Alderson, Krahnke, & Stansfield, 1987). At the time, no English language development standards existed to serve as outcome goals for ELL students or to provide guidance in test development, so the tests were developed with different theories and understandings about what constituted English proficiency.

Following this landmark case, in 1981, the Fifth Circuit Court established a three-part test to evaluate the adequacy of a district's program for ELLs in the case of Castañeda v. Pickard. The court required districts to answer three questions in relation to district programs for ELLs: (a) Is the program based on an educational theory recognized as sound by some experts in the field or is considered by experts as a legitimate experimental strategy; (b) Are the programs and practices, including resources and personnel, reasonably calculated to implement this theory effectively; and (c) Does the school district evaluate its programs and make adjustments where

needed to ensure language barriers are actually being overcome? These questions currently

guide school districts in developing and maintaining ELL programs. The reauthorization of the

Elementary and Secondary Education Act (ESEA)—renamed the No Child Left Behind (NCLB)

legislation in the United States in 2001—requires that all students, even if they do not speak

English, be tested in English annually for academic progress (Gándara & Baca, 2008). This

federal law applies to all states receiving federal monies for education.

The first ESEA legislation was passed in 1965, with an overall purpose of improving

educational opportunities for poor children. At that time, President Lyndon B. Johnson, a former

teacher who had witnessed poverty's impact on his students, believed that equal access to

education was vital to a child's ability to lead a productive life. To date, this legislation is still

the most expansive federal education bill ever passed (Landsberg, 2004).

In 1994, another federal legislative act was passed, known as the Improving America's

Schools Act (IASA). This legislation was enacted to provide federal financial aid to states for

the implementation of standards-based reform. The act set the foundation for requiring states to

implement accountability systems as an extension of standards-based reform (Franklin, 2011).

The significance of these legislative acts is the accountability or monitoring of students

including ELLs built into the legislation. The legislation requires some type of testing that is

usually summative in nature, as it "provides useful information . . . of students' achievement or

progress at the end of a course of study" (Bachman & Palmer, 1996, p. 98) or a school year. The

required testing measures student progress in relation to a set of adopted standards.

In summary, although the federal government took important steps to support schools in

providing an equal educational opportunity for all students, student achievement has only

slightly improved. According to the Education Trust Equity Alert of 2010, student academic achievement has improved in ELA and math by 2 percentage points, but subgroup performance (ELLs are a subgroup), still shows that too many students are performing below grade level (*Student Achievement in California,* 2010).

State Legislation

If they accept the fiscal resources that accompany legislation, states are responsible for implementing the federal mandates. With the enactment of NCLB, states were required to create English language proficiency (ELP) tests to monitor student progress. As of 2007, the ELP tests states used to monitor progress had been available for the previous 15 to 20 years. The focus of these ELP tests was on social or general uses of language rather than on the language of the classroom, textbooks, educational standards, or content-area assessments (Bailey, 2007). Consequently, there was a gap between what was being tested with the ELP tests and what was required for ELLs to keep up academically.

Prior to 2002 and before NCLB, California passed two state mandates that changed education for ELLs. The first was Proposition 227, a California state initiative designed to "reduce or eliminate the use of languages other than English for classroom instruction" (Parrish, Linquanti, & Merickel, 2002). Before the passage of Proposition 227, 30% (California Department of Education, 2009) of ELLs participated in a bilingual program that provided instruction in their primary language. Proposition 227 explicitly called for "sheltered English immersion during a temporary transition period not normally intended to exceed one year" (Hakuta, Butler, & Witt, 2000). Whereas Prop 227 suggested that ELLs could learn English in one year and perform academically on the same level as their peers, this claim "is totally without

empirical foundation" (Cummins, 2011, p. 143). In fact, this notion of accelerated English

proficiency via English immersion instruction is "contradicted by the research literature"

(Goldenberg, 2008, p. 12).

The second mandate, Assembly Bill 748, approved by the State Board of Education

(SBE) in 1999, mandated the creation of the English Language Development (ELD) Standards

and assessment. The SBE was required to identify or develop a test to assess the language

proficiency levels of English language learners, an effort that resulted in the California English

Language Development Test (CELDT). The ELD standards provided English proficiency

outcomes, and the CELDT was aligned to these standards. The CELDT has been revised nearly

every year since 2001, raising issues with its reliability and validity (Stokes-Guinan &

Goldenberg, 2011).

In summary, federal and state legislation define the parameters of the educational system,

which is directly tied to standards and standardized testing. Standardized testing has become a

foundation in United States education, and tests are used to place, assess, remove, and admit

students to different instructional programs (Nieto, 2004). Such testing is meant to scrutinize the

education system and hold Local Education Agencies (LEA) accountable for the education being

provided to students; however, researchers have found that "tests correlate more with family

income than with intelligence or ability, and the result is that poor students of all backgrounds

are unfairly jeopardized in the process" (Nieto, 2004, p. 406). Furthermore, many of these

students are still learning the English language.

History of Standardized Testing

According to Grodsky, Warren, and Felts (2008), standardized testing in the United States began following the use of intelligence testing to group students by ability. Intelligence testing was first used to identify students in need of special education services but expanded to classifying students for instructional purposes. In the 1890s, the president of Harvard used ability grouping to improve instructional efficiency. By the early 1900s, most elementary schools and many urban high schools grouped students by ability, as indicated on intelligence tests. The first large-scale group test of ability was developed by psychologists for the US Army in 1917–1918. These psychologists subscribed to the position that one could be quite intelligent but illiterate or not proficient in the English language (Fulcher, 2010; Sticht & Armstrong, n.d.).

Standardized educational testing continued to expand as advances were made in statistics and measurement. At the same time, immigration increased, creating more diverse school populations and a need to improve instruction, thus amplifying the demand for testing. As ability tests continued to gain popularity, achievement tests were being created. The differences between these tests were that ability tests were measures of cognitive ability, whereas achievement tests were measures of curricular content. The first voluntary state-wide achievement test was the Iowa Test of Educational Development, developed by E. F. Linquist in 1929 (Grodsky et al., 2008).

Although standardized testing has provided a way to assess student achievement and thus evaluate how well schools are doing, a persistent achievement gap separates minority students and nonminority students. According to the 2011 National Assessment of Educational Progress (NAEP), no change occurred in reading achievement between these two groups in grade four but

improvement was indicated in grade eight. Over the period of 1998–2011, the gap narrowed

between 8th-grade White and Black students in the State of Delaware. In the same time period,

gaps in White and Hispanic scores narrowed in only two states: Oregon and California (Institute

of Educational Sciences, 2012). In California, although student performance had risen in almost

every subject and grade level, a substantial achievement gap persisted between low-income and

higher-income students and between African American and Latino students and their White and

Asian peers (Frey, 2012). Latino students make up the majority of the ELL population

(Solorzano, 2008). These gaps in achievement become a critical issue for education in the U.S.

particularly because "they bode ill for English learners' future educational and vocational

options. They also bode ill for society as a whole, since the costs of large-scale

underachievement are very high" (Goldenberg, 2008, p. 11).

Standardized testing has become the foremost method for measuring students' academic

success and has had multiple affects on their educational experience. However, according to

Nieto (2004), an evaluation of standardized tests conducted by the FairTest organization found

that "testing programs in most states need a complete overhaul, or at least major improvements,

to actually achieve what they claim to be doing" (p. 98). Further, all tests have limitations; for

example, "Any test that employs language is, in part, a measure of their language skills. This is

of particular concern for test takers whose first language is not the language of the test" (AERA

et al., 1999, p. 91). To understand these limitations, educators need to be knowledgeable about

testing theories and test types.

Testing Theories

Since the enactment of No Child Left Behind in the U.S., standardized testing has become paramount in educational accountability. NCLB is intended to close the achievement gap between minorities and nonminorities, with assessment being a key component in this legislative mandate (Spinelli, 2008). Educators must be knowledgeable about achievement measures and their results—especially for subgroups of students that need more assistance. Large-scale testing most often utilizes two key testing theories.

Test theory refers to the procedures for estimating key characteristics of a test or measure, such as validity and reliability, dictated by the measurement model, which has a unique set of assumptions and is based on a statistical or mathematical model (Suen, 1990, as cited in Zhu et al., 2011). The two main theories used in educational testing are classical test theory and item response theory, which are most widely used for accountability. These theories provide the foundation for creating different types of tests.

Classical Test Theory

Classical test theory (CTT) has served as the foundation for measurement theory for several decades. Classical tests are based on how an individual performed on a test in comparison to how others performed on the same test (Mason, 2007). In CTT, "the true score of a person can be found by taking the mean score that a person would get on the same test if they had an infinite number of testing sessions" (Kline, 2005, p. 91). Mean score is the average score. Because collecting an infinite number of test scores is impossible, a true score can only ever be hypothetical. Although classical theory has been used for many years, its greatest limitation is in

controlling for the random error score, which are unknown and unpredictable changes in the experiment that prohibit obtaining a true score. Less error represents a truer score.

Item Response Theory

The most current testing theory being used in high stakes accountability tests is item response theory (IRT), which focuses on responses to individual items—in contrast to the sum of the scores privileged by classical theory. IRT assumes that each examinee has some underlying ability—also known as a *latent trait*—that relates to the probability of answering items correctly. This latent trait is unobservable, so it must be placed on a scale of measurement. In IRT, the scale of difficulty is very easy, easy, medium, hard, and very hard. The discrimination levels are none, low, moderate, high, and perfect. All items are identified on these two levels of difficulty and discrimination (Baker, 2001). In IRT, the primary purpose for administering a test to an examinee is to locate that person on the ability scale in order to evaluate how much underlying ability he or she possesses. In addition, comparisons among examinees can be made for purposes of assigning grades, awarding scholarships, and so forth (Baker, 2001).

Briefly, classical test theory focuses on the examinee's raw test score—the sum of the scores received on the items in the test. IRT focuses on whether the examinee got each item correct or incorrect (Baker, 2001). Regardless of which test theory is employed, language factors may be a source of measurement error when testing ELLs due to unnecessary linguistic complexity (Abedi, 2002). Knowledge of the types of tests administered to students is essential to understanding test results and how those results can be used for making decisions about students.

Types of Tests

Norm-referenced tests. To be meaningful, test scores have to be associated with some type of reference point. In the case of norm-referenced tests (NRT), scores are used "to locate an examinee on the distribution of scores of all examinees" (Grodsky et al., 2008). Scores that are norm-referenced are compared to a group of people that already took the test—known as the "norming" group." The comparison of scores is between the test taker and his/her peers rather than a set of standards. In California, the assessment system known as Standardized Testing and Reporting (STAR) has a norm-referenced component that is administered only in grades three and seven through the 2011 school year (http://www.cde.ca.gov/ta/tg/sr/cefstar.asp). Consequently, ELLs are administered the test in English, and the norming group is made up of English speakers, not English language learners.

Criterion-referenced tests. Criterion-referenced tests are designed to determine whether students have mastered specific content such as the knowledge of content standards. An example of this type of test is the California Standards Tests (CST), which all students in grades 2-11 must take and is administered and written in English. The CST is currently used in the content areas of English language arts, mathematics, science, and history–social science. To reiterate, ELLs are given this test in English to assess their knowledge of the respective content areas that they have been learning in English. The California English Language Development Test (CELDT) is a criterion-referenced test aligned to the English Language Development (ELD) standards, which assess the English language proficiency of pupils whose primary language is a language other than English.

These two types of tests—norm-referenced and criterion-referenced—are well researched in the literature, and several researchers have found them to be inappropriate for testing ELLs who have the dual challenge of learning academic content and learning English (Abedi, 2002; Murphy, Bailey, & Butler, 2006; Stokes-Guinan & Goldenberg, 2011; Wolf, Kao, Bachman, Bailey, & Farnsworth, 2008). Apprehensions about these test types focus on the reliability and validity of the instruments. According to the *Standards for Educational and Psychological Testing* (AERA et al., 1999), all tests used in educational and psychological testing must be reliable and valid There are acknowledged methods of establishing validity and reliability for tests used in education.

Validity Theory

Validity theory answers two fundamental questions in testing: (a) What interpretations can be made regarding a student's performance on a test; and (b) What decisions can be made based on the results? *Validity* refers to the degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed purpose (AERA et al., 1999, p. 11). Validation begins with a proposed interpretation of test scores referring to a particular construct. According to Bachman and Palmer (1996), a construct is "the specific definition of an ability that provides the bases for a given test" (p. 21). In other words, the characteristic or concept that a test is designed to measure is the construct (American Educational Research Association et al., 1999). Interpretations or inferences are made after the test is administered, and that information is used to make decisions about the test taker. Interpretations must be validated or justified for those decisions.

Several researchers have discussed validity and test use in assessment in attempts to merge the two concepts. For example, Bachman and Palmer (1996) posited a model of test usefulness specific to language testing. This model explains the characteristics of language inherent in testing. Test usefulness is viewed as a unifying principle that embodies other important principles of test construction and evaluation in relation to language. Their view consists of six interrelated principles: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Without reliability, all other test qualities are irrelevant (Abedi, 2007). Therefore, Bachman and Palmer (1996) described reliability first—with methods for determining reliability—then followed with brief descriptions of the five additional principles of test usefulness.

Reliability

Reliability refers to the consistency of the instrument. Reliability is the degree to which scores on a test are consistent over repeated administrations with the same population of individuals or groups (AERA et al., 1999). A more formal definition is "the degree to which a test consistently measures whatever it measures" (Gay, Mills, & Airasian, 2009).

According to the *Standards for Educational and Psychological Testing* (AERA et al., 1999) (hereafter referred to as *Standards*), several methods can determine the reliability of a test: test-retest, parallel forms, internal consistency, and inter-rater reliability. The test-retest method requires a test be given to the same group of people on two occasions. Then the two sets of scores are correlated. The most critical factor in test-retest is the amount of time between test administrations, assuming the construct does not change over that period. A shorter time gap

provides a higher correlation, whereas a longer time gap has a lower correlation. The statistical analysis for this method is Pearson r because the variables are continuous (Gay et al., 2009).

With the parallel forms method, two tests are created that have parallel test items. A large number of items that measure the construct is required to create two forms of the test. Both forms would be administered to the same group of people; then the scores would be correlated, providing an estimate of the reliability. As with the test-retest method, Pearson r is the appropriate statistical analysis for this method.

Internal consistency refers to how well someone performs on subsets of the same test. According to Mason (2007), "Internal consistency procedures may be based how half of the test correlates with the other equivalent half or how items of the test correlate with each other" (p. 30). The statistical analyses for internal consistency are Cronbach's alpha (α) or Kuder-Richardson (KR-20), depending on the sample size and type of data. Coefficient alpha models for dichotomous data such as Cronbach's α are equivalent to the KR-20 coefficient with the difference being that Cronbach can handle both dichotomous and continuous variables.

Inter-rater reliability refers to how two people rate the same tester using the same test. If a test is consistent, two raters should rate the same tester similarly. Again a high correlation between raters would reveal more consistency—versus a low correlation indicating low consistency. Pearson r statistical analysis would be applied in this case as well.

In summary, the methods for determining reliability ensure that the test consistently measures the same construct over repeated administrations with the same individuals or groups. Goldenberg (2011) defined reliability as the degree to which a measure is consistent and

dependably produces the same result. Without reliability, validity cannot be measured. For this reason, educators must ensure that any test used with students is reliable.

The next principle in Bachman and Palmer's (1996) test usefulness model is construct validity, which refers to the degree to which a test accurately measures what it says it measures. Therefore the construct being measured must be clearly defined.

The principle of authenticity refers to the relevance of the test task to the target language use domain. In other words, authenticity refers to whether the test items or tasks relate to the type of language indicative of a real world situation. Interactiveness as a principle has to do with how the test engages a test taker's language ability. Interactiveness is defined as the extent to which the test tasks engage the processes and strategies that are part of the construct being assessed (Vleuten & Elsevier, 2007). The construct in this model is language. The principle of impact refers to test use at macro and micro levels. The test has a direct impact on the test taker, the test user, and the entire education system because decisions are made from the test results. Practicality involves test implementation and resources. If a test is too costly to administer and score, it will not be used (Bachman, 2009). In sum, Bachman and Palmer's (1996) model of test usefulness brought specific characteristics of language testing to the forefront.

Validity

A seminal author in educational measurement, Samuel Messick (1989) defined validity as "an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (p. 33; as cited in Hubley & Zumbo, 2011). Messick's definition of validity contributed to the creation of the *Standards* developed by three organizations to guide the development and use of tests. The

Standards are intended for individuals and organizations that develop, administer, and use test results to make inferences about individuals or groups. The three contributing organizations are the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (AERA et al., 1999). Five documents preceded the current 1999 edition, which is presently under revision. Specifically, the *Standards* were developed to "promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (AERA et al., 1999, p. 1).

The *Standards* provide criteria for the "evaluation of tests, testing practices, and the effects of test use" (AERA et al., 1999, p. 2). According to Abedi (2007):

> First, validity is a property of the inferences (i.e., the interpretations) that one draws from test scores, rather than being a property of the test or a property of the test score. Second, validity is never proven or established, but is argued on the basis of an ever-accumulating body of evidence that speaks to the accuracy of the inferences that one makes on the basis of test scores. (p. 20)

Distinct types of validity identified in traditional theory include criterion validity, content validity, and construct validity. The *Standards* refer to sources of validity *evidence* rather than distinct types because validity is defined as the "degree to which all the accumulated <u>evidence</u> supports the intended interpretation of test scores for the proposed purpose" (AERA et al., 1999, p. 11). The *Standards* enumerate five sources of validity evidence.

First, validity evidence based on test content refers to the themes, wording, and format of the items, tasks, or procedures regarding test administration and scoring. Content validity evidence refers to the relationship of the test content to the construct being measured (AERA et al., 1999). One way of determining this type of evidence can be through alignment studies.

Studies have been conducted to align English Language Proficiency (ELP) tests to English

Language Development (ELD) standards—for example, the CELDT.

Validity evidence based on test content can also be determined through a sensitivity

review. In a sensitivity review of ELP tests, a group of experts familiar with the language and

culture of ELL students is convened to determine if the test material is objectionable in any way

to ELLs. The review would include consideration of whether the material is offensive to ELLs,

portrays ELLs unfavorably or stereotypically, would pose an advantage or disadvantage for

ELLs, or is unfamiliar to ELLs. The goal of a sensitivity review is to identify items, passages,

and other test material that should reviewed for appropriateness from a content perspective

(Sireci, Han, & Wells, 2008).

Second, validity evidence based on the internal structure of a test is determined by

analyzing the relationships of test items and test components and whether these conform to the

construct on which the score interpretations are based. For example, differential item

functioning (DIF) can be used to screen test items for bias. DIF occurs when different groups of

test takers who have the same total test score have different average item scores or, in some

cases, different rates of various item options (AERA et al., 1999). A content review committee

analyzes the information and, in some cases, removes items from the test.

Next, validity evidence based on response processes is an analysis of the test takers'

cognitive processes and strategies for responding to test items and the degree of congruency to

the construct being measured (Sireci et al., 2008). For example, if a test is measuring

mathematic reasoning, examinees should be reasoning about the given problem and not about

using an algorithm (AERA et al., 1999). Unfortunately, according to Sireci et al. (2008), few

studies have been conducted in this area with respect to ELLs; however, this research is imperative to confirming that ELLs and non-ELLs use the same cognitive processes and strategies when responding to test items.

Validity evidence based on relations to other variables was traditionally known as criterion-related validity. It involves correlating test scores to other measures of the same construct to identify convergence or discrimination (Sireci et al., 2008). The *Standards* (1999) provide the example that if a student takes a multiple-choice reading comprehension test, his or her performance should relate to other measures of reading comprehension to establish convergence. If the reading comprehension scores do not highly correlate with a mathematics test then there is discrimination.

Another type of evidence in this category is predictive validity, which establishes the accuracy with which test scores predict performance on the same construct at a later date. This type of validity evidence was measured in the current study between ADEPT and CELDT.

Reliability and validity are the cornerstones of ethical testing, yet many of the aforementioned methods have not been used in the development of ELP tests. Baily has also found that recent ELP tests have been used for multiple purposes yet haven't been validated as such which is not recommended by the *Standards* (Herman, Bachman, & Bailey, 2008; Sireci et al., 2008; Wolf et al., 2008).

A broader issue with testing has to do with power structures in our society. The NCLB mandate initiated by the federal government to make schools and districts accountable for educating all students brought transparency to the subgroups of students that were being left behind; yet it did not provide a solution to the problem. In fact, the mandate was underfunded,

leaving school districts with a budget shortfall that, in fact, prevented its full implementation. Shohamy (2001) has found ample evidence that tests are "powerful tools, often introduced in undemocratic and unethical ways for disciplinary purposes and for carrying out various policy agendas" (p. 2). Additionally, Shohamy (2001) has demonstrated how groups in authority use tests to manipulate educational systems and to impose their personal agendas. The high stakes environment of educational testing increases the misuse of tests and even cheating. Yet tests still perform as powerful instruments of manipulation, because they serve as an "unwritten contract between those in power, who want to dominate, and those who are subject to tests, who want to be dominated in order to maintain their place and status in society" (Shohamy, 2001, p. 375). In California, high stakes tests in education result in public scrutiny about schools due to the Academic Performance Index (API), which measures the academic performance and growth of schools on a variety of academic measures, including the achievement of ELLs in academic content and English proficiency.

Although principles are in place for the ethical use of tests and test creation, the larger problem of test use must be considered in language testing. According to Shohamy (2001), "Use-oriented testing, is concerned with the uses of tests in their relation to curriculum, ethicality, social class, politics and knowledge, and their impact on individuals and educational systems" (p. 374). To better understand how the principles of validity and reliability impact the use of language testing and how individuals learn a second language in addition to their first language, a review of second language acquisition theories is necessary.

Second Language Acquisition Theory

The concept of a "common underlying proficiency" is essential to this component of the theoretical framework because, as Cummins briefly stated, in the course of learning one language, a child acquires a set of skills and implicit metalinguistic knowledge that can be drawn upon when working in another language. In other words, first and second language proficiency are related, and students can learn academic content in their first language, which will transfer to the second language. In short, learning conceptual knowledge in the second language that was learned in the first language is not necessary. If the student acquired the conceptual knowledge adequately then he or she merely needs to learn the label in the second language. Therefore, being proficient in a first language supports acquisition of a second language (Second Language Acquisition, 1996).

Another contribution that Cummins made to the understanding of language acquisition is in articulating the difference between conversational, everyday language and academic language, also known as *the language of school.* Cummins labeled these two types of language Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP). BICS is acquired more quickly than CALP, and ELLs must receive instruction that is comprehensible in order to develop CALP. The language of the classroom, CALP, tends to be abstract, whereas BICS can be acquired more easily because it is accompanied by gestures and other clues that support comprehension. These two distinctions of language are essential to understanding second language acquisition theories and English language proficiency (ELP) tests (Second Language Acquisition, 1996).

By definition, second language acquisition is the process by which humans learn a language in addition to their "first" or native language (Abedi, 2007). Many theories have been developed to explain the process of second language acquisition and have produced accompanying instructional methods and practices. According to Conteh-Morgan (2002), the three main categories or theories are *behaviorism, innatism,* and *interactionism.* A more recent resource of second language theories from Lavadenz (2011) expands on Conteh-Morgan (2002) with four language theories: (a) structural, (b) cognitive, (c) functional, and (d) interactional. An understanding of these approaches and theories is critical to understanding English language instruction and testing.

Behaviorism, first proposed by B. F. Skinner, explains language development as influenced by environmental stimuli, whereby association, reinforcement, and imitation are the primary factors in the acquisition of language. Alternately, innatist theories attribute to humans the natural ability to process linguistic rules (Abedi, 2007). Innatism or cognitive language theory emerged from the idea that humans are born with the ability to create and use language independent of experience (Conteh-Morgan, 2002; Lavadenz, 2011). Children detect the rules of grammar in language as they learn in a natural way; they do not merely imitate the sounds they hear. Noam Chomsky (1957) labeled this internal ability the "language acquisition device" (LAD) (as cited in Lavadenz, 2011, p. 21). According to Chomsky, the LAD remains active during a critical language-learning period, generally understood as childhood, and then turns off, making it more difficult to learn language as an adult.

Krashen and Terrell (1983) proposed a model of second language acquisition in the category of innatist theory. Five hypotheses emerged from Krashen's observations of Terrell's

teaching practices with second language learners. The natural approach to learning language, as proposed by Krashen and Terrell (1983), follows a linear developmental sequence with a focus on communication supported by meaningful language input. The five hypotheses are: (a) Humans are genetically programmed to learn language, and explicit instruction may interfere with that process; (b) The natural learning of language proceeds from simple to complex; (c) The affective filter allows one to experiment with language and take risks if kept to a minimal level; (d) Comprehensible input means that communication must be delivered in a way that the learner understands and that is just above the learner's proficiency level; (e) The language monitor focuses on language rules and correctness but can inhibit the production of language Kindergarten through 12th-grade (K–12) classrooms have used these hypotheses extensively (Lavadenz, 2011).

Interactionism or *interactional language theory* is focused more on the use of language in communicative acts, on the functions of language, and its use in various contexts. This theory supports the idea that humans have innate language abilities, but also asserts that language is learned through social interactions, which is a tenet of sociocultural theory. Interactionists believe that native speakers modify their language when they communicate with language learners in order to accommodate the learners' communicative proficiency and level of understanding (Conteh-Morgan, 2002).

Functional language theories view language as the medium for achieving specific purposes or conveying specific meanings. Communication—not just the grammar and structure of a language—forms the essential characteristic of language (Lavadenz, 2011). This theory also states that form serves function in language. Finally, structural theories equate language with its

linguistic features, such as the phonological, lexical, and syntactical components. Language

learning in this theory requires knowledge of linguistic forms (Lavadenz, 2011). These different

theories of language inform the testing and instructional practices that schools use today.

Second Language Acquisition Tests

When NCLB was implemented in 2001, some states did not have a state-wide assessment

system in place—much less a high-quality method for testing their rapidly growing ELL

populations (National Education Association, 2008). Also, states were not required to have ELP

standards, so English language development instruction consisted of a "locally devised

curriculum, which could range from exceptional to seriously inadequate" (Albers, Kenyon, &

Boals, 2008, p. 75). Thus, ELP tests differed across states.

Two resources in the literature cite five of the most common English language

proficiency tests. Although the resources are 10 years apart, Esquinca, Yaden, and Rueda (2005)

determined the five most common tests from a 2000–2001 survey administered prior to the

implementation of NCLB. A comparison of these two resources resulted in four most common

tests: the Basic Inventory of Natural Language (BINL), the Idea Proficiency Test (IPT),

Language Assessment Scales (LAS), and Woodcock-Muñoz Language Survey (Esquinca et al.,

2005; Vecchio & Guerrero, 1995). Table 1 shows these four tests and includes summaries of the

validity and reliability of each test. Four additional widely used but not most common ELP tests

are summarized in Appendix A.

Table 1

Most Common English Language Proficiency Tests

Test	Grades/Ages	Purpose	Domains Measured
Basic Inventory of Natural Language (BINL)	K–12	To be used as an indicator of oral language dominance and/or language proficiency	Speaking
Idea Oral Language Proficiency Test (IPT)	PreK–12	To assess overall language proficiency	Listening, speaking, reading, and writing
Language Assessment Scales (LAS)	K–adult	To determine a level of oral language proficiency and literacy	Listening and speaking
Woodcock-Muñoz Language Survey	PreK–college	To measure student's cognitive academic language proficiency	Listening, speaking, reading, writing

Adapted from "Current Language Proficiency Tests and Their Implications for Preschool English Language Learners" by A. Esquinca, D. Yaden, & R.t Rueda, *ISB4: Proceedings of the 4th International Symposium on Bilingualism*, p. 676. Copyright 2005 by the University of Southern California.

Most of these tests share similar strengths and weaknesses in reliability and validity. Whereas the reliability coefficients of BINL and IPT are high, there are shortcomings in validity. The reliability of the BINL was determined through split-half correlation, yielding a correlation of .925. Test-retest analyses were also conducted, yielding a correlation coefficient of .80 (Vecchio & Guerrero, 1995). The reliability of the IPT Oral Proficiency tests is extremely high: .99 based on interitem correlations. Internal consistency for the entire IPT reading battery was .76, whereas the correlations for all items of the writing samples ranging from a low of .90 to a high of .98. The writing sample correlations were determined through inter-rater reliability. The LAS-Oral claims to have high reliability overall, but the listening subtest coefficients were much lower.

All tests claim to have some form of validity. For example, BINL and IPT claim to have content and construct validity—but BINL also claims to be predictive, whereas IPT claims to have criterion validity. The LAS-Oral Technical Manual provides statistical evidence to support the validity of the oral portions of the test.

The only test that claims to measure Cognitive Academic Language Proficiency is the Woodcock-Muñoz Language Survey. Although it claims to have extensive evidence for reliability and validity, the Woodcock-Muñoz Language Survey only provides details in its administration manual, which is not readily available, making it difficult to compare to other ELP tests.

Despite problems in methodology, all the abovementioned ELP tests report differing levels of reliability and validity. Two of the tests measure all domains of language, which are listening, speaking, reading and writing, whereas another only assesses two language domains. One of the tests only measures the domain of speaking.

The aforementioned ELP tests were designed prior to NCLB, so there was no requirement that they be standards-based. At that time, most states had requirements for identifying and assessing ELLs but the choice of ELP test was up to the local district. Therefore, comparisons of performance among schools, districts, or states were a challenge. In addition, the ELP tests were created from a different conceptual model, which focused on social or conversational skills not academic language skills, thus hindering the ability to use the results diagnostically for instruction.

Although Vecchio and Guerrero (1995) and Esquinca et al. (2005) have provided important and useful information regarding the validity of these ELP tests and the domains

measured, the greatest challenge to date with language testing is the absence of a generally accepted definition of language proficiency (Stokes-Guinan & Goldenberg, 2011). To reiterate, language is a complex construct. Therefore the theories of language acquisition assist in understanding the language skills that can be measured. Consequently, when a test of language is validated, it is understood that language skills are being measured and other aspects of language may not have been included in the test. For example, if a test is validated for structural purposes of language, such as language's phonological, lexical and syntactical features, it may not measure how well meaning is conveyed. The validity of a test is only as useful as our understanding of the construct. For this reason, all ELP tests need to be used with caution and in conjunction with other measures of language, such as formative assessments.

Classroom Assessments of ELP

A body of research examines formative assessment that teachers can use to determine ELL progress in English language proficiency. These types of assessments—also known as classroom measures—range from informal reading inventories, structured observations, end-of-unit tests, and performance tasks, such as writing samples, speeches, demonstrations, student portfolios, interviews, conferences, and student self-assessments (Gottlieb, 2006; O'Malley & Valdez Pierce, 1996) A common characteristic of formative assessments is that they are teacher created or designed in collaboration with classroom teachers. Often these types of assessments do not meet the rigor of standardized tests (which require validation), yet they provide a more comprehensive view of students' abilities.

An example of a commonly used formative assessment is the Structured Oral Language Observation Matrix (SOLOM) developed by the San Jose Area Bilingual Consortium in

California. SOLOM is a rating scale used to assess oral proficiency only. The SOLOM is not a test, but rather a rating scale that teachers can use to assess their students' command of oral language (Hargett, 1998). The most obvious limitation to the SOLOM is the subjectivity of the raters, who are subject to many influences that affect the ratings they give. Hence, SOLOM is best used in conjunction with other measures of language proficiency.

Another standards-based test of English proficiency is the English Language Development (ELD) Classroom assessment. This assessment documents a student's progress by level and domain toward each of the California ELD standards. The California English language development standards state five proficiency levels and three domains: listening/speaking, reading, and writing. Using this structure, the (ELD) Classroom assessment consists of all of the standards for each level and domain. Teachers have to score student's progress toward each standard, using the following scale:

4 = Advanced Progress: Exceeds the standards for the identified ELD level.

3 = Average Progress: Meets the standards for the identified ELD level.

2 = Partial Progress: Demonstrates some progress toward mastery of the standards.

1 = Limited Progress: Demonstrates little or no progress toward mastery of the standards.

Advancement to the next ELD level requires that a student receive scores of 3 or 4 on all of the standards. When a student masters all the standards in level 5, reclassification can be considered. Similar to A Developmental English Proficiency Test (ADEPT), this assessment is intended to inform teachers and guide their instruction in meeting students' needs (Llosa, 2008).

Finally, with several researchers expressing the need for more formative assessment, a grant recently funded by the Carnegie Corporation was given to the Wisconsin Center for

Education Research. This 3-year grant proposes to develop a formative assessment model for ELLs in the middle and high school grades using best practices. The intent is to develop valid and reliable formative measures of ELP, improve the learning and achievement of ELLs, and provide teachers with practical tools for keeping ELLs on track for academic success (http://flareassessment.org/).

These descriptions provide a glance of the types of ELP tests that were available before the federal mandate for ELD standards and assessments. Of the three classroom assessments, the first is not considered a test, the second was designed with the ELD standards, and the third has not been completed but attempts to validate formative measures of ELP. This area of formative ELP assessment is growing, thus there is hope that more studies will be published in the near future. Although some ELP tests have been redesigned to meet the requirements of NCLB, challenges to meeting accountability mandates and to following the Standards for Educational and Psychological Testing persist.

Concerns with ELP Tests Designed for NCLB Accountability

Before 2001, ELLs were often excluded from accountability systems that required testing in English. In California, many ELLs were in bilingual programs and tested in their primary language. However, with the No Child Left Behind Act 2001 (NCLB), ELLs must progress in acquiring English and demonstrate yearly progress in math and reading.

NCLB requires states to find or create a test of English language proficiency to measure progress in acquiring English. Some states select commercially produced tests whereas others participate in consortiums to collaborate on creating this test. According to Bailey (2007), in 2007, many states were using "the older generation of assessments that focus on social or general

uses of language" (p. 2), whereas others were still responding to the mandate. The problem with the older assessments is that they were not aligned with the current standards, curriculum, and textbooks. In addition, there was a lack of information about the language demands of school and the length of time it takes for an ELL to achieve proficiency (Bailey, 2007; Hakuta et al., 2000). Further, Bailey (2008) found that many states were using their ELP test for multiple purposes, such as diagnosis, placement, and identification. One test cannot serve multiple purposes, as the "professional standards require evidence of validity for each intended use of any assessment" (Herman et al., 2008; Wolf et al., 2008), which was not the case for many of the common ELP tests. Post-NCLB consortium-developed tests were created for the purpose of evaluation. Thus far, the tests are valid for this purpose, and states are able to calculate student progress. Unfortunately, what NCLB did not address was the need for formative assessments to guide classroom instruction, which teachers need in addition to the evaluative tests (Bunch, 2011).

In summary, due to accountability requirements, assessment of ELLs in the U.S. is directly influenced by federal and state legislation. In the haste to have accountability measures in place, psychometric principles of test construction—including test use and interpretation—have been overlooked (Solorzano, 2008). In spite of unambiguous findings in the research literature and recommendations by researchers, ELLs are expected to become proficient in English in one year, which conflicts with current second language acquisition research (Bailey, 2007; Goldenberg & Coleman, 2010; Hakuta et al., 2000). In addition, several of the ELP tests reviewed only measure certain domains of language and only provide summative information. There is yet to be developed a comprehensive ELP test that addresses all theories of language

48

and all domains of language—both social and academic language—and that has been validated

for the multiple purposes education needs.

Theoretical Framework

When conceptualizing the testing of English Language learners in education, testing theory must

be taken into consideration. Testing theory seeks to explicate how students are tested for English

language proficiency (ELP). ELP is a function of second language acquisition, which includes

theories that explain how a second language is learned in addition to a first language. The

theoretical framework for this study would not be complete without sociocultural theory (SCT),

also a learning theory. Ultimately, students must be proficient in the primary language so they

can access the academic content on which they are being tested. SCT completes the context in

which an ELL experiences the educational system. (See Figure 1).
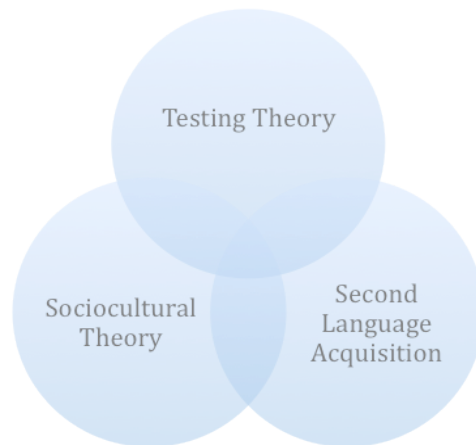
Figure 1. Conceptualizing testing of ELLs for English proficiency.

Sociocultural Theory

Sociocultural views of learning include the premise that second language teaching and

learning take place through the social interactions of learners and more capable others and seek

to understand cultural and historical influences on learning (Faltis & Hudelson, 1998; Lave & Wenger, 1991; Wertsch, 1991, as cited in Lavadenz, 2011). Lev Vygotsky, who lived from 1896–1934, developed this framework while another researcher was developing what is known today as cognitive development.

Walqui (2006) has described the main tenets of Vygotsky's theory as follows:

- Learning precedes development.

- Language is the main vehicle (tool) of thought.

- Mediation is central to learning.

- Social interaction is the basis of learning and development. Learning is a process of apprenticeship and internalisation in which skills and knowledge are transformed from the social into the cognitive plane.

- The Zone of Proximal Development (ZPD) is the primary activity space in which learning occurs.

Although all of these tenets are important for this study, the role of language in learning is most relevant.

Vygotsky described two critical roles that language plays in cognitive development. First, language constitutes the main means through which adults convey information to children. Language develops through social interactions for the purpose of communication. Once it is internalized, language becomes thought. Therefore, thought results from the development of language (McLeod, 2007). The second critical role of language is that it becomes a very powerful tool for intellectual adaptation (McLeod, 2007). Culture shapes a person's learning and internalization of language. The culture in which a language is learned influences language

development. Vygotsky believed that children are born with the basic abilities for intellectual development. As they participate in social interactions in cultural contexts, children develop more advanced abilities, referred to as the "tools of intellectual adaptation"—or how to think. Because they are culturally determined, these tools can vary. As such, language is a product of the culture within which it is learned (Lantolf & Thorne, 2002).

Another important tenet of Vygotsky's theory is that learning occurs through mediation, which is "the idea that humans rely upon tools and other social and cultural artifacts to regulate the world around them" (Cross, 2010, p. 440). One of these tools is language. While learning a language, children observe language use in others and then replicate it in speech, and eventually in thought. One form of mediation in language learning is regulation, which occurs in three phases. The first phase is object regulation, in which objects in the environment regulate a child's thinking. An illustrative example is when students need objects to solve mathematical problems—such as using beans to complete addition problems. The second phase is other regulation, which is when the student has assistance from peers and/or adults. This type of assistance is referred to as *scaffolding*, which is employed when the learner is in the Zone of Proximal Development (ZPD). The ZPD refers to the space in which we learn with support from others. The last phase of regulation is self-regulation, which refers to the stage in which activities are accomplished with little to no assistance (Lantolf, 1996). The importance of the last phase is to understand that one must be able to self-regulate based upon the context of communication. For example, in challenging situations, or when meaning is compromised for some reason, a proficient speaker may need to revert to previous phases such as other or object regulation to maintain communication. As Lantolf and Thorne (2002) have stated, "Language is

the most pervasive and powerful cultural artifact that humans possess to mediate their connection

to the world, to each other and to themselves."

Finally, the ZPD can be understood in relation to Vygotsky's concept of the More

Knowledgeable Other (MKO). This critical understanding of how language is acquired

underscores the importance of social interaction. The principle of the ZPD conceptualizes the

difference between what one can do independently and what one can do with guidance. The

MKO serves as the guide from that which is not known to that which is known. The ZPD

constitutes the space in which we can learn from the guidance and support of the MKO and

move to the next level of cognitive development (McLeod, 2007)

Simultaneous to Vygotsky's theory of development, a psychologist names Jean Piaget

established cognitive development. Although cognitive development does not have an all-

encompassing explanation, Piaget's theory provides a different perspective of learning and its

influence on language learning. Unlike Vygotsky, Piaget believed that learning occurred in

universal stages and in a particular sequence. He also placed greater emphasis on the role of

children's play in learning and believed that development is necessary and precedes learning,

including language learning. Vygotsky placed more emphasis on social factors, cultural

influences, and the role of language in cognitive development (McLeod, 2007).

Sociocultural theory completes the theoretical framework for this study due to its

emphasis on social mediation and meaningful interactions between experts and novices as these

concepts relate to oral proficiency testing in general (Brooks, 2009). Lantolf and Thorne (2007)

have stated, "SCT is grounded in a perspective that does not separate the individual from the

social and in fact argues that the individual emerges from social interaction and as such is always

fundamentally a social being" (p. 217–218). As Walqui (2006) stated with regard to learning

language, "The cognitive and the social go hand in hand," and "Education never takes place in a

vacuum but is deeply embedded in a sociocultural milieu" (p. 159).

Summary

The achievement of English language learners has received a great deal of attention in

recent years due to the passing of NCLB, which requires districts to evaluate the academic

achievement of students by disaggregating data, which consists of student test scores (Public

Law 107 – 110 107th Congress An Act, 2002). This charge has brought to the forefront a long-

standing gap in achievement among subgroups of students including ELLs. The ELL subgroup's

unique characteristic is the challenge it faces in learning English in addition to maintaining

academic achievement. The time it takes for an ELL to become proficient in English far exceeds

the time allowed for yearly academic achievement (Goldenberg & Coleman, 2010; Hakuta et al.,

2000).

Although substantial literature assesses test development and creation, studies specific to

language testing seem to have been neglected in recent years due to the urgent need for

accountability measures. Several researchers have noted the problem of testing ELLs in English

if they are not proficient in the language (Abedi, 2002; Bailey, 2007; Bailey & Huang, 2011;

Gándara & Baca, 2008; Goldenberg & Coleman, 2010; Hakuta et al., 2000). The Standards for

Educational and Psychological Testing provide criteria for testing, but has not changed the

current trend of requiring ELLs to become proficient in English and to attain academic

achievement comparable to their English-speaking peers. When students are not fully proficient

in the language of a given test, the results may be compromised because they do not accurately reflect the competencies the test intended to measure (AERA et al., 1999).

To address the need for formative measures of English language proficiency, this study aimed to validate ADEPT, which can be used up to three times during the school year when teachers can use the results to modify instruction. With ADEPT results, teachers obtain specific information about the grammatical structures ELLs struggle with at different proficiency levels that can preclude them from attaining English proficiency.

CHAPTER 3

METHODOLOGY

Introduction

The purpose of this study was to establish the reliability and validity of an oral language assessment titled A Developmental English Proficiency Test (ADEPT). ADEPT only tests the skills of listening and speaking in English, therefore these two sections were compared to the listening and speaking subsections of the California English Language Development Test (CELDT) in order to establish concurrent validity. ADEPT is used to monitor English language learner (ELL) progress toward English proficiency during the school year and can be used with all ELLs two to three times per year.

This study was quantitative in nature using archival student data provided by a California public school district. Student data obtained from the district were anonymous. No direct contact occurred between the researcher and the students. This chapter includes the research design, research questions, setting, data, data collection, procedures, and data analysis techniques employed in the study.

Research Design

A correlational research design using archival data was used to compare ADEPT listening and speaking scores to CELDT listening and speaking subscores. The rationale for choosing a correlational design was to describe an existing condition. More specifically, "Correlational research involves collecting data to determine whether, and to what degree, a relationship exists between two or more quantifiable variables" (Gay et al., 2009, p.196).

To address the research questions, four years of student scores on CELDT and ADEPT were collected in Winter 2013, with permission from Highland School District. Using multiple years of data allowed the researcher to look at the pattern of reliability and validity of ADEPT over time. The district gave each student an anonymous identification number in order to match performance on the two tests being compared.

Research Questions

The research questions investigated were:

RQ1. What is the reliability of ADEPT?

RQ2. What is the overall concurrent validity of ADEPT and CELDT in the listening and speaking subscores?

RQ3. How well does the student's ADEPT score predict the subsequent CELDT?

The first question was addressed by using descriptive statistics to analyze the internal consistency of ADEPT. The statistical analysis the researcher employed was Cronbach's alpha (α), which is equivalent to the KR-20 coefficient, according to the Statistical Package for the Social Sciences (SPSS) system. The difference between these two measures is that Cronbach's α can handle both dichotomous and continuous variables. The researcher used Pearson r analysis to address the second research question regarding concurrent validity. The final question was measured using Pearson r as well. All statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS) system.

Setting

Due to the need to select a sample of ELL students who take both ADEPT and CELDT, the researcher used nonrandom purposive sampling (Gay et al., 2009). The selected district,

Highland School District, was located in Los Angeles County in the Southern California basin

adjacent to the Los Angeles International Airport. The district was established in 1907.

Highland School District educated close to 10,000 pre-kindergarten through 12th grade students

in a diverse urban community. At the time of this study, there were seven elementary schools,

three middle schools, and a charter high school. Students who did not attend the charter high

school attended a high school in a neighboring district. The district has a significant percentage

of ELLs in its student population. Table 2 shows the number of ELLs by level, by school, and

by grade level. The distribution of ELLs is concentrated at the K–5 elementary level. The

middle school ELL population is approximately 20% of the total ELL population.

Table 2

Number of ELLs by School and by Grade Level 2011–2012

Elementary Schools	Total No. by School	Grade K	1	2	3	4	5
School Euc	526	100	103	96	84	60	83
School Jeff	246	44	39	50	42	36	35
School Korn	325	78	48	56	58	42	43
School Ram	357	68	61	74	55	52	47
School Wash	337	68	65	82	54	34	34
School York	327	65	64	64	54	45	35
School Zel Dav	488	103	108	101	77	56	43
Total Elementary	2606						
Middle Schools		Grade 6	7	8			
School Bud Car	175	86	48	41			
School Haw Mid	198	81	59	58			
School Prai Vis	190	85	53	52			
Total Middle	563						
High Schools		Grade 9	10	11	12		
School Math/Science Acad	16	8	6	1	1		
Total High	XX						
Total in District	3186						

Note. Adapted from DataQuest, developed and maintained by the California Department of Education,
http://dq.cde.ca.gov/dataquest

According to the California Department of Education, Educational Demographics Unit, the total school population in the 2011–2012 school year was 8,866 students. (See Table 3) The district served a racially and ethnically diverse population with a large proportion of children participating in the free and reduced lunch. The district had a substantial population of Hispanic/Latino students, and about one fifth of the population was Black or African American. The number of English language learners is significant to this study.

Table 3

Demographic Characteristics of the District, 2011–2012 School Year

Special Programs	No. of students	% of Enrollment
English Learners	3,186	35.9
Free/Reduced Price Meals	5,863	88.3
Compensatory Education	8,848	99.9
Race/Ethnicity		
American Indian or Alaska Native	40	0.5
Asian	259	2.9
Native Hawaiian or Pacific Islander	129	1.5
Filipino	158	1.8
Hispanic or Latino	6,148	69.3
Black or African American	1,912	21.6
White	188	2.1
Two or More Races	28	0.3
Not Reported	4	0.0
Total number of students	8866	

Note. Adapted from Ed-Data, www.ed-data.k12.ca.us/

Data

The district determined that ADEPT would replace the Student Oral Language Observation Matrix (SOLOM) that was being used for ELLs who scored at the beginning or early intermediate level on CELDT. The district required the use of SOLOM each trimester to monitor progress, but felt it was ineffective and too subjective. The annual CELDT data was too old to be effective in informing instruction as the window closes in October, which is after school had begun. Therefore, in exploring other tests to replace SOLOM, the district selected

ADEPT for use with all kindergarten through 8th-grade English language learners. (See Table 4)

Also, ADEPT would be used with ELLs who were at the beginning or early intermediate level of

proficiency based on CELDT scores. (See Table 5) To analyze the pattern of reliability and

validity of the ADEPT over time, this study used multiple years of data.

Table 4

Total Number of ELLs in the District by School Year and Grade Level

Grade	Number of ELLs		
	2009/10	2010/11	2011/12
Kindergarten	525	469	526
1	494	545	488
2	495	481	523
3	520	456	424
4	416	420	325
5	358	380	321
6	301	263	252
7	244	190	160
8	235	164	151
9	17	7	8
10	8	7	6
11	4	4	1
12	2	0	1
Total by School Year	3619	3386	3,186

Note. Adapted from DataQuest, developed and maintained by the California Department of Education, http://dq.cde.ca.gov/dataquest.

Of the total ELL population, only students at the beginning and early intermediate level

on CELDT were tested with ADEPT at each trimester. Based on the district criteria, about 12%

of the 2011–2012 ELL population was assessed with ADEPT. The results from the students who

took CELDT and ADEPT yearly were analyzed by the researcher.

Table 5

Total Number of Students Scoring at the Beginning and Early Intermediate Level on CELDT

Grade	2009–2010	2010–2011	2011–2012	2012–2013
1	75	91	94	94
2	73	94	125	82
3	71	76	65	94
4	41	57	39	39
5	20	29	35	33
6	1	23	32	42
7	0	1	7	38
8	1	1	0	11
Total	282	372	397	433

Note. Adapted from Highland School District, DataDirector [TM], Riverside Publishing, 2010

The reliability study was based on four years of ADEPT results data. Concurrent and predictive validity was based on three school years of data from CELDT and ADEPT. CELDT measures language proficiency by testing the four language domains of listening, speaking, reading, and writing by grade clusters, resulting in students being categorized into one of the five performance levels: beginning, early intermediate, intermediate, early advanced, or advanced. ADEPT only measures the domains of listening and speaking in four performance levels: beginning, early intermediate, intermediate, and early advanced. Therefore, only the subscale scores from the listening and speaking domains were compared. Descriptions of each test are provided. Sample test items are in Appendix C.

The California English Language Development Test (CELDT)

CELDT was developed in response to legislation requiring school districts to annually assess the English language proficiency of all students with a primary language other than English. All students in transitional kindergarten through 12th grade whose primary language is not English based on responses to a home language survey are tested within 30 days of enrollment

or 60 days prior to instruction for initial English language proficiency identification with the CELDT.

The CELDT is a criterion-referenced test aligned to the State Board of Education's (SBE) adopted English Language Development (ELD) Standards, which assess the English language proficiency of pupils whose primary language is not English. CELDT contains items that measure how proficient students are in the English language. The items cover the four domains of language: listening, speaking, reading, and writing in English. The CELDT does not measure achievement on the California academic subject frameworks and standards (*California English Language Development Test Technical Report 2010–11*, 2011).

The CELDT was developed by the California Department of Education Statewide Assessment Division (*California English Language Development Test Technical Report 2010–11*, 2011). The assessment is administered once a year until a student reaches proficiency in the four domains of language: listening, speaking, reading and writing. An overall score is comprised of the composite of these four domains. The comprehension score is a composite of the reading and listening domains. Five separate tests are grouped into the following grade clusters: kindergarten and grade one, grade two, grades three through five, grades six through eight, and grades nine through 12.

The CELDT contains three basic item formats, the first of which is multiple choice. Multiple-choice items consist of a question and three or four response choices. The second format is dichotomous-constructed response, which requires that the student generate a verbal response that is recorded as correct or incorrect. The third format consists of constructed-response items that are evaluated by comparison to a rubric and scored on a point scale from 0

through 4.

The CELDT has been revised every year since the field-testing in 2000. The original scale and performance cut scores were created from the 2000 field-test and the first edition in 2001–2002. In 2006, the test was rescaled, establishing new performance cut scores. In 2009–2010 the domains of reading and writing were including in K–1, and a standard setting was conducted in January 2010 to establish performance cuts scores.

According to the CELDT technical report (2010–2011), the validity of the test is an ongoing process:

> Although we have no external measures available at present to correlate with the CELDT scale scores, the pattern of correlations within the CELDT provides preliminary validity evidence by showing that the correlations among the four language domains are positive and reasonably high. (California English Language Development Test Technical Report 2010–11, 2011, p. 73)

The CELDT has four item formats for listening and speaking described in Table 6.

Table 6

CELDT Item Formats for Listening/Speaking Domains

Domain	Format
Listening	*Following Oral Directions:* Items require students to identify classroom-related nouns, verbs, and prepositions, and to demonstrate understanding of the relationships of words without having to read or reconfigure the directions to show aural comprehension.
	Teacher Talk: Items require students to comprehend important details, make high-level summaries, and understand classroom directions and common contexts.
	Extended Listening Comprehension: Items require students to follow the thread of a story, dialogue, and/or presentation of ideas, extract more details, pick out what is important, and use inference, and listen to learn.
	Rhyming: Items require students to demonstrate aural discrimination of medial and final sounds in English words by producing a word that rhymes with a pair of rhyming words presented by the examiner (grades K–1 and 2 only).
Speaking	*Oral Vocabulary:* Items elicit a single word or short phrase, and assess simple to complex social, academic, and classroom vocabulary.
	Speech Functions: Items elicit one declarative or interrogative statement; assess formation of a response appropriate to a situation; and focus on question formation.
	Choose and Give Reasons: Items elicit two sentences or complete thoughts, and assess independent clause formation and the ability to make rudimentary explanations or persuasive statements.
	Comprehension: Items require students to identify basic text features such as book titles.

Note. Adapted from California English Language Development Test Technical Report 2009-10, California Department of Education, 2011

CELDT Administration Procedures

According to the 2012–2013 CELDT Information Guide, the testing windows for the 2012–2013 year are July 1 through October 31 for the Annual Assessment, and July 1 through June 30 for the Initial Assessment. The CELDT is an untimed test. For students in transitional kindergarten, kindergarten, and grade one, listening, reading, and writing domains are administered individually, and the estimated time required is approximately 15 to 30 minutes per domain. For students in grades two through 12, the listening, reading, and writing domains are administered as a group and take about two hours to complete. The speaking part of the test is administered individually to all students in transitional kindergarten through grade 12 and takes about 10 to 15 minutes for each student to complete. Only test examiners who are employees of the Local Education Agency, are proficient in speaking English, and have received formal CELDT training may administer the test (California Department of Education, 2012).

CELDT Scoring and Interpretation

The three formats of the CELDT are multiple-choice, dichotomous-constructed response, and constructed-response. The first two formats elicit responses that are recorded on scannable documents and then machined scored. The constructed-response items are also scanned but then scored by the contractor's scorers. Constructed-response items are associated with the writing and speaking domains and are graded by human readers. Many districts train their own readers for the constructed-response scoring. All scores are then merged using a pre-identification number for each student tested and performance levels are determined.

When interpreting students scores on the CELDT, one must note that in kindergarten and first grade, the overall score is calculated as 45% listening, 45% speaking, 5% reading, and 5% writing. In grades 2-12, the overall scores is the average of the four domains. On the Student

Performance Level Report, the CELDT reports scores as "scale scores" expressed as three digit numbers ranging from 140 to 810. Lower numbers signify less proficiency, whereas higher numbers indicate more proficiency. Scale scores are reported for each of the four domains as well as for overall proficiency. Comprehension is an average of the reading and listening scale scores. Once performance levels are determined, the results are provided to teachers, parents, administrators, and the California Department of Education for state and accountability purposes.

Districts and schools may use CELDT results to make decisions about program placement for ELLs, as one form of exit criteria from an ELL program and to gauge English proficiency progress in EL programs. However, CELDT results should not be used as the single indicator for making these decisions because the performance levels are very broad and must be used with caution. Hence, multiple measures must be used in the decision-making process.

A Developmental English Proficiency Test (ADEPT)

ADEPT is a classroom-based oral language test of the listening and speaking domains of English. Four levels of English proficiency are tested: beginning, early intermediate, intermediate, and early advanced. Items require either a nonverbal response, such as pointing, gesturing, or a dichotomous-constructed response, generating an oral response to a question or a prompt. ADEPT is designed to assess student responses to strategically worded questions or prompts. Student responses reveal key structures or forms that are within the students' command, providing insight to the teacher regarding level of proficiency.

The first edition of ADEPT was created by a school district in Northern California to monitor student progress in oral English proficiency and was used for more than 20 years. In the 2000–2001school year, the California Reading and Literature Project (CRLP) made adaptations

to the test, which were field-tested that year by hundreds of teachers in California. The 2001

edition of ADEPT was aligned with 2001–2002 edition of CELDT. In 2004, a pilot study to

establish the concurrent validity of the ADEPT to CELDT was conducted by the University of

California Educational Partnership Center at University of California, Santa Cruz. The results of

the ADEPT total scores were correlated to the scaled CELDT scores in listening and speaking,

yielding a correlation of .76. This result indicated good concurrent validity between the two

tests, suggesting that ADEPT captures a good portion of the listening and speaking skills tested

by the CELDT. The 2006 edition of ADEPT was used for this study.

The ADEPT assessment measures two domains of language; listening and speaking,

identified as receptive and expressive items, respectively. Table 7 shows language skills assessed

at each proficiency level.

Table 7

ADEPT Language Skills in the Listening/Speaking Domains by Proficiency Level

Domain	Proficiency Level
Receptive Level 1 Expressive Level 1	*Beginning*- Ability to understand basic vocabulary and generate one or two word responses and one sentence with the present progressive.
Receptive Level 2 Expressive Level 2	*Early Intermediate*- Ability to understand and generate routine expressions and utterances using the present progressive, is/are, and pronouns.
Receptive Level 3 Expressive Level 3	*Intermediate*- Ability to understand and generate utterances with varied verb tenses, possessives, pronouns, and adverbs.
Expressive Level 4	*Early Advanced*- Ability to use more complex pronouns, adverbs, and varied verb forms.

Note. Adapted from A Development English Proficiency Test Assessment Manual, California Reading and Literature Project, 2006

ADEPT Administration Procedures

The test administration guide states that ADEPT can be used to assess all English learners two or three times per year. Because districts in California are required to administer the CELDT during the annual assessment window from July 1 through October 31, many choose to administer ADEPT in the winter to monitor progress, and again in spring to measure growth.

To determine what level at which to begin testing a student, the teacher (or whoever is administering the test) must consider if the student is speaking in phrases and/or sentences. If this is true, testing should begin at the early intermediate level or with the corresponding CELDT level. If students are not producing phrases, testing should commence at the beginning level.

The materials needed to administer ADEPT include:

1. ADEPT assessment manual and CD

2. Copies of Student Scoring Sheets

3. Clipboard and pencil

4. Consumable copies of illustrations at the beginning level

5. Pencil and crayons at the beginning level

The setting for administering ADEPT should be a relatively quiet space so that the tester can easily hear the student's responses. The seating arrangement should be such that the student only sees the illustrations in the manual, and the tester sees the prompt pages. The student score sheets should be kept on a clipboard for easy recording and not in the student's view. When beginning the assessment, the initial instructions to the student are:

> We are going to talk about some pictures in English to give you a chance to show how much English you know. There are several different ways to answer correctly. Don't worry if you don't know the answer. This is to help me find out more about what you know so I can help you learn even more English. I will be writing parts of what you say

so I can remember later, so I need to be able to hear exactly what you say. If I don't quite understand you, I may ask you to repeat. This doesn't mean it's wrong, just that I didn't hear clearly. So just repeat exactly what you said the first time. If you don't hear something I have said, or if you want me to repeat it so you can hear it again, please ask me to say it again and I will be glad to do so. (p. 6)

The tester is required to speak only in English, using a normal speaking voice, enunciating clearly but with natural inflection. The instructions and wording on the tester's prompt page must be followed exactly, including directions of where to point on the picture when indicated. The student should be aware that his or her responses are being recorded or written down. All responses must be accepted neutrally or with a positive comment even if the response is incorrect. Allowing the student several seconds of wait time is allowed; rushing the student or pushing to move onto the next level should be avoided. Once the student responds, the next item must be prompted without delay to avoid inefficient testing procedures, which invite distraction or increased testing time. Student responses must be recorded promptly to ensure accurate scoring. It is important to capture language that is an accurate representation of what the student can produce naturally.

When scoring student responses, 0 indicates an incorrect response or no response, and 1 indicates a correct response. Recording the phrase containing the target structure exactly as it is stated—whether correct or incorrect—is also important. Any mispronunciation due to a second language "accent" are not counted as errors. A word, phrase, or complete sentence can constitute a student response. Responses are dependent on the prompt. Responses are considered utterances and must be grammatically correct in order to be scored as 1. There are two cases in which a response is incorrect even if the utterance makes sense. First, if a response includes the correct target structure but the entire utterance is grammatically incorrect, the response is scored

68

as incorrect. Second, if a response includes an incorrect target structure but the entire utterance is grammatically correct, the response is scored as incorrect. In either case, the response is incorrect.

Each test level has a benchmark indicating the number of correct responses required before moving on to the next level. The beginning level (B) has 48 receptive items and 10 expressive items. The combined benchmark is 48 out of 58. The early intermediate level (EI) has 23 receptive items with a benchmark of 18, which must be met before moving on to the EI expressive level. The EI expressive level has 13 items, with a benchmark of 10. The intermediate level (I) has 18 receptive items, with a benchmark of 14. The intermediate expressive level has 11 items, with a benchmark of 9. The advanced level does not have any receptive items. It only has 10 expressive items, with a benchmark of 8. The student must meet each benchmark before moving to the next level.

Item formats in CELDT and ADEPT

The CELDT contains three basic item formats: multiple-choice (MC), dichotomous-constructed-response (DCR), and constructed-response (CR). (See Table 8) CELDT multiple-choice items consist of a stem (question) and three or four response options. Dichotomous-constructed-response items, which are found primarily in the speaking test, usually require a constructed response (i.e., a reply to a question), which is then evaluated as right or wrong by the test examiner. Constructed-response items are evaluated with respect to a rubric and may receive 0 through 4 points (California English Language Development Test Technical Report 2010–11, 2011).

Table 8

Number of CELDT Operational Items in the Listening/Speaking Domains by Grade Cluster

Grade-Level Cluster	Domain	No. of Items	No. of Items by Type		
			DCR	MC	CR
K–1	Listening	20	10	10	0
	Speaking	20	13	0	7
2	Listening	20	10	10	0
	Speaking	20	13	0	7
3-5	Listening	20	0	20	0
	Speaking	20	13	0	7
6–8	Listening	20	0	20	0
	Speaking	20	13	0	7

Note. DCR =Dichotomous Constructed Response, MC =Multiple Choice, CR = Constructed Response

ADEPT consists of seven subtests in four proficiency levels: beginning level 1, early intermediate level 2, intermediate level 3, and early advanced level 4. Levels 1–3 assess receptive and expressive skills, but level 4 only assesses expressive skills. Receptive skills relate to the domain of listening whereas expressive skills relate to the domain of speaking. (See Table 9) ADEPT has two item formats, nonverbal response and dichotomous-constructed response, which was compared to CELDT, which itself has three item formats. The researcher was not able to obtain an item analysis of CELDT by language domain, so the subscale scores in listening and speaking were compared to the overall scores in the ADEPT receptive and expressive levels.

Table 9

Number of ADEPT Operational Items in the Listening/Speaking Domains

Proficiency	Domain	No. of Items	No. of DCR Items
Beginning	Receptive Level 1	48	0
	Expressive Level 1	10	10
Early Intermediate	Receptive Level 2	23	0
	Expressive Level 2	13	13
Intermediate	Receptive Level 3	18	0
	Expressive Level 3	11	11
Early Advanced	Expressive Level 4	10	10

Note. DCR indicates dichotomous constructed response

Although the test formats for the listening and speaking domains differ in ADEPT and CELDT, an overlap exists in the following CELDT formats: following oral directions, oral vocabulary, and speech functions. (See Appendix C for item examples.) These items require a dichotomous constructed response, which is similar to the expressive items in ADEPT. Table 10 shows the ADEPT score ranges for each level with the cutscore, also known as the benchmark. Table 11 shows the CELDT initial/annual scale score ranges.

Table 10

ADEPT Score Ranges

No. of Level	Level of Test	No. of Items	Cutscore
1	Beginning Receptive	1–48	Combined w/Expressive score
1	Beginning Expressive	1–10	48
2	Early Intermediate Receptive	1–23	18
2	Early Intermediate Expressive	1–13	10
3	Intermediate Receptive	1–18	14
3	Intermediate Expressive	1–11	9
4	Early Advanced Expressive	1–10	8

Table 11

CELDT Initial/Annual Scale Score Ranges

		1	2	3	4	5	6	7	8
Beginning	Listening	220–361	220-374	220-388	220-401	220-410	230–412	230–417	230–426
	Speaking	140–352	140–369	200–387	200–404	200–410	225–416	225–422	225–422
Early Intermediate	Listening	362–408	375–425	389–442	402–460	411–472	413–483	418–494	427–507
	Speaking	353–404	370–419	388–435	405–450	411–458	417–466	423–475	423–479
Intermediate	Listening	409–454	426–475	443–497	461–518	473–536	484–569	495–571	508–594
	Speaking	405–456	420–469	436–481	451–496	459–506	467–517	476–527	480–538
Early Advanced	Listening	455–501	476–526	498–551	519–577	537–600	570–637	572–648	595–669
	Speaking	457–508	470–519	482–531	497–542	507–555	518–567	528–580	539–594

Data Collection

The researcher gained access to Highland District in July 2011 by contacting the newly appointed superintendent. The researcher had an existing relationship with the superintendent, having provided professional development in the district over the past several years. After an initial meeting with the superintendent, the researcher was directed to communicate with the

district English learner (EL) specialist, who would work with the district data manager to retrieve the data necessary for the study.

The researcher had been in contact with the district English learner specialist via email from July 2011 until the present to determine what ELL data were available in their data system. The researcher created and sent Excel spreadsheets to the EL specialist, illustrating what data were needed and in what format. (See Tables 12 and 13) There were few challenges obtaining the necessary data for this study.

Table 12

K–8 ADEPT *and* CELDT *Scores by School Year for Students at the Beginning and Early Intermediate Level*

Student Identifier	Celdt Scores 2009 5 Scores[a]	Adept Score Nov 2009 2 Scores[b]	Adept Score Mar 2010 2 Scores	Celdt Scores 2010 5 Scores	Adept Score Nov 2010 2 Scores	Adept Score Mar 2011[c] 2 Scores
Student A						
Student B						

a. 5 scores refer to listening, speaking, reading, writing, and overall. *b.* ADEPT 2 scores refer to listening and speaking
c. 2011–2012 and 2012–2013 school year columns not shown due to space limitations

Table 13

Data Collection Spreadsheet to Calculate Internal Consistency in ADEPT *Level 1*

Student Identifier				Item No.							CONTINUE WITH LEVEL 2
	R1.1	R1.2	R1.3	R.16	R1.7	**ETC**	E1.1	E1.2	E.1.3	E1.4	
Student A											
Student B											

Note. Spreadsheet created by researcher to illustrate data needed to run reliability analyses.

All data regarding student CELDT and ADEPT scores were obtained by permission from the school district from the internal data system called DataDirector[TM]. All ELLs in grades K–8 who score at the beginning or early intermediate level on CELDT are given the ADEPT test in

November (Fall) and March (Spring) of the corresponding school year. These students' subscale scores comprised the data set. No student names were associated with the data. All student information was completely anonymous. The primary data set was the results of the Fall (November) and Spring (March) ADEPT tests for the 2009–2010, 2010–2011, 2011–12, 2012–13 school years. The secondary dataset was Highland District's annual CELDT test scores from the same school years. All data were analyzed using the Statistical Program for the Social Sciences (SPSS), a statistical processing software program available to LMU students.

Procedures

The researcher received the first data file for correlation analysis from the district in February 2013, and four additional files for the reliability study in March 2013. All files were in Microsoft Excel spreadsheets. The following paragraphs detail the process and decision criteria used to prepare the files for statistical analysis.

Because the data were retrieved from a district's data system, the researcher had to make two assumptions. First, it is assumed that the teachers who administer ADEPT follow the administration guidelines recommended in the administration manual. Second, it is assumed that the district data system is properly set up to record and retrieve ADEPT and CELDT data.

The Excel file retrieved for correlation analysis included 392 records for the school years covering 2009–2013. The first step in preparing the Excel file for analysis was to remove any student records that did not have data or did not have CELDT or ADEPT scores for any school year. To determine concurrent and predictive validity, each record had to have scores for CELDT and ADEPT.

The column headers were formatted to wrap the text in order to see the titles of each

column. The Excel file had all the corresponding columns listed in Table 8, but ADEPT Level 1

columns included each subtopic not an overall score. For example, Level 1 receptive had seven

columns corresponding to the seven subtopics of the test. Level 1 expressive had two columns

representing the subtopics. Therefore, a column was added to the right of the receptive topic

columns with a formula to calculate the sum, resulting in a total receptive score for Level 1. This

column represents an index of the student's performance at Level 1 receptive. The same

procedure was followed to obtain a total expressive score or index for Level 1 expressive. (See

Table 14) Once the sum columns were calculated, the subtopic columns were hidden.

Table 14

ADEPT Columns with Summary Column

Student Identifier	[ADEPT 1st Test: Beginning Level 1 - 2009–2010] Receptive - Family and Clothing[a] Score	[ADEPT 1st Test: Beginning Level 1 - 2009–2010] Receptive - Food Score	[ADEPT 1st Test: Beginning Level 1 - 2009–2010] Receptive - Animals, Size and Number Score	2009–10 ADEPT Fall Test Beginning Level 1 Total Score Receptive	[ADEPT 1st Test: Beginning Level 1 - 2009–2010] Expressive - Survival Language Score	[ADEPT 1st Test: Beginning Level 1 - 2009–2010] Expressive - Present Progressive Score	2009–10 Benchmark 48–58 With Rec ADEPT Fall Test Beginning Level1 Total Score Expressive
1.	1	0	5	14	0	0	0

Note. Example of columns as they were received from the district's DataDirector[TM] system.
[a]There are 7 topic columns for Level 1 receptive and 2 topic columns for Level 1 expressive. Only 3 are shown here in receptive due to
space limitations.

Next, another column was added for the ADEPT Instructional Level, representing the

level at which the student did not meet the benchmark. Testing ceases at the level where the

student does not meet the benchmark indicating this is the students instructional level. Student

scores in the receptive and expressive columns of levels 1–3 were compared to the benchmarks

to determine the ADEPT instructional level, and that number was entered into the overall level

column, as shown in Table 15.

Table 15

ADEPT Levels with Overall Columns Only for Beginning Level 1

Student Identifier	2012–13 ADEPT FALL Test Beginning Level 1 Total Score Receptive	2012–13 ADEPT FALL Test Beginning Level1 Total Score Expressive	2012–13 ADEPT FALL Test Early Intermediate Level 2 Receptive Score	2012–13 ADEPT FALL Test Early Intermediate Level 2 Expressive Score	2012–13 ADEPT FALL Test Intermediate Level 3 Receptive Score	2012–13 ADEPT FALL Test Intermediate Level 3 Expressive Score	**12–13 ADEPT RECEPTIVE Overall**	**12–13 ADEPT EXPRESSIVE Overall**
1.	47[a]	10	22	13	14	8	**3**	**3**

Note. The bolded columns indicate the ADEPT overall level based on the comparison to the benchmarks of each level 1–3. Benchmarks are found in the administration guide.

[a] These are sample scores for one student in ADEPT levels 1–3 listening and speaking

77

Finally, the columns for levels 1–3 were hidden in order to reveal only CELDT scores and ADEPT instructional levels for each year. There were columns for ADEPT in Spring of each year (except 2009–2010 and 2012–2013), but these columns were also hidden because they would not be used in the statistical calculations. This decision was based on protocol that states that the second administration assesses only missed items, not the entire battery again. Therefore, only CELDT annual listening and speaking scores and ADEPT Fall receptive and expressive scores were used in the analysis, as shown in Table 16.

Table 16

CELDT Levels with ADEPT Overall Levels

Student Identifier	12–13 CELDT Proficiency Level Listening	12–13 CELDT Proficiency Level Speaking	12–13 CELDT Proficiency Level Overall Test	12–13 ADEPT RECEPTIVE Overall	12–13 ADEPT EXPRESSIVE Overall
1.	2	3	2	2	2

Note. Columns not needed for analyses are hidden, as compared to Table 13.

The Excel files for internal consistency analysis were received by school and class, so several Excel files were merged into one. (See Table 17.) It was not necessary to remove any records.

Table 17

Reliability Data from One Class

Question Name	R3.1	R3.2	R3.3	R3.4	R3.5[a]
Question Numbering	1	2	3	4	5
Answer Key	Y	Y	Y	Y	Y
Student 1	Y*	N	Y*	N	Y*
2	Y*	Y*	Y*	Y*	Y*
Percent Correct	100%	50%	100%	50%	100%

Note. Student names were removed and replaced with student and number. Asterisks were obtained in original source.
[a]Only five questions are shown for illustration. There are 29 items total in this level.

After all files were merged, the fields with Y* were converted into 1, and the blank fields were converted into 0. There were some fields with N, meaning "no response," but they were converted into 0 as well. For statistical purposes, fields must be numeric. The resulting Excel file for ADEPT Level 1 is illustrated in Table 18.

Table 18

Data for Internal Consistency Analysis ADEPT Level 1

Question Name	R1.1	R1.2	R1.3	R1.4	R1.5	R1.6
Question Numbering	1	2	3	4	5	6
Answer Key	Y	Y	Y	Y	Y	Y
Student 1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	1	1	0	1	0	0
5	1	0	0	0	0	0

Note. Resulting spreadsheet once letters and blank spaces were converted to 0 or 1. Answer key did not have to be converted for statistical analyses to be performed. Percent correct row deleted.

Data Analysis

Statistical analysis was performed using Statistical Program for the Social Sciences (SPSS). Institutional Review Board (IRB) approval was requested and approved in Spring 2013. Descriptive statistics were used to analyze the psychometric properties of ADEPT. Internal consistency was estimated using Cronbach's alpha to determine if all the items were measuring the same construct. Although, Kuder-Richardson 20 could have been used in the analyses, Cronbach's α is more commonly used because the results would be equivalent according to SPSS. Initially, Pearson r analysis was used to determine concurrent validity, but Spearman's Rho was also calculated due to some small sample sizes ($n<50$). Concurrent validity is designed to measure how well a particular test correlates to a previously validated test (Gay et al., 2009). Predictive validity was measured using Pearson r and Spearman's Rho analysis, but the results

were similar so only Pearson *r* results are reported in Chapter Four.

The researcher has included a discussion of the results in Chapter Four based on these data analyses. In the future, these results will also be provided to the district.

Summary

This chapter has provided an overview of the research design, research questions, setting, data, data collection, and data analysis for this study. The goal of this study is to confirm that ADEPT is a valid and reliable assessment.

CHAPTER 4

RESULTS AND ANALYSIS

Introduction

The purpose of this study was to determine the reliability and validity of A

Developmental English Proficiency Test (ADEPT). This classroom measure of oral language

tests the skills of listening and speaking. It is used to monitor English language learner (ELL)

progress toward English proficiency during the school year. Using ADEPT, along with the

California English Language Development Test (CELDT), exceeds what is required by

California law, as explained thusly:

> California law requires students in kindergarten through grade twelve whose home
> language is not English to take an English skills test. This test helps schools identify
> students who need to improve their skills in listening, speaking, reading, and writing in
> English. Schools also give the test each year to students who are still learning English.
> (CDE, n.d.)

The CELDT is an annual test, whereas ADEPT is used 2 to 3 times per year and not

required by California law. Therefore, it is critical to document the reliability and validity of

ADEPT and its correlation to the current CELDT in order to employ this test with confidence.

This chapter describes the statistical analyses performed and the results obtained related

to each research question. Each research question is presented, followed by a summary and a

table representing the results.

Research Questions

RQ1. What is the reliability of ADEPT?

Reliability refers to the degree to which test scores for a group of test takers are

consistent over repeated applications of a measurement procedure, and hence are inferred to be

dependable and repeatable for an individual test taker (AERA et al., 1999). Tests must first be

reliable in order to be valid.

To determine the reliability of ADEPT, the researcher obtained Highland School District

results from the Fall testing period of 2012. Four Excel files were created, representing the four

levels of ADEPT. The Excel files included every student tested and every test item response in

each level for the Fall testing period.

There are a total of 133 ADEPT items, 89 items that measure receptive or listening skills

and 44 that measure expressive or speaking skills. The domains of listening and speaking were

analyzed for each level of ADEPT. The reliability analysis measured the extent to which these

items are consistent among themselves and with the test as a whole (Gay et al., 2009). The

reliability of each ADEPT level was calculated using Cronbach's α index of internal consistency.

Although, Kuder-Richardson 20 (KR-20) could have been used in the analyses, it is analogous to

Cronbach's α, which is also used for nondichotomous measures. The Cronbach's alpha scores

are presented in Table 19.

Table 19

Internal Consistency Coefficients for ADEPT Levels 1–3

Level	Title	Cronbach's Alpha	N	Mean	Variance	SD	No. of Items
1	Beginning Receptive	.97	110	36.81	158.36	12.58	48
	Beginning Expressive	.87	110	8.01	7.170	2.68	10
2	Early Intermediate Receptive	.84	297	20.24	11.11	3.33	23
	Early Intermediate Expressive	.87	297	7.30	15.73	3.40	13
3	Intermediate Receptive	<u>.53</u>	65	13.12	3.90	1.97	16[a]
	Intermediate Expressive	.82	65	4.72	10.1	3.17	11
4	Early Advanced Expressive	-	2[b]	-	-	-	10

Note.[a]Total number of items is 18. Two items were dropped due to the mean score of 1.00. Underlined
score is considered low. [b]Not enough data to analyze.

For ADEPT Level 1, Cronbach's α internal consistency coefficient for the 48-item receptive domain revealed a correlation of $\alpha = .97$, $n = 110$. The 10-item expressive domain in the same level resulted in a correlation of $\alpha = .87$, $n = 113$ indicating a good level of internal consistency for both domains of level 1.

ADEPT Level 2 items were analyzed separately for internal consistency. Cronbach's α coefficient for the 23-item receptive domain was $\alpha = .84$, $n = 300$. The 13-item expressive domain yielded a correlation of $\alpha = .87$, $n = 300$. Both domains yielded a good level of internal consistency.

ADEPT Level 3 items were analyzed for internal consistency, resulting in $\alpha = .53$, $n = 65$ for the 16-item receptive domain and $\alpha = .82$, $n = 65$ for the 11-item expressive domain. Receptive item numbers 6 and 10 were dropped from the calculation because the mean for each was 1.00. The receptive domain of level 3 had poor internal consistency compared to the expressive domain, which had good internal consistency.

Only 2 student responses were obtained for ADEPT Level 4, so the analysis was not performed. All item statistics for each level, receptive and expressive are in Appendix D.

In summary, the first three ADEPT levels showed internal consistency with Cronbach's alpha scores, ranging from a low of .53 to a high of .97, with a median of .86. The first criterion of validity is reliability, so concurrent validity and predictive validity were explored with the next two research questions.

RQ2. What is the overall concurrent validity of ADEPT and CELDT in the listening and speaking subscores?

Concurrent validity refers to "the degree to which the scores on a test are related to the scores on a similar test administered in the same time frame" (Gay et al., 2009). Because the CELDT annual test must be completed by October 31 of the school year, and the first ADEPT administration is in November (considered the Fall administration), concurrent validity was measured with the data from Highland School District. To measure concurrent validity, the CELDT listening and speaking subscale scores were compared to the Fall ADEPT listening and speaking scores of each school year. (See Table 20)

Table 20

2012–13 August CELDT Subscale Scores in Listening/Speaking with November ADEPT Receptive/Expressive Scores

Student	12-13 AUG CELDT Proficiency Level Listening	12-13 AUG CELDT Proficiency Level Speaking	12-13 NOV ADEPT Receptive overall	12-13 NOV ADEPT Expressive Overall
1	2	3	2	2

Note. Although 4 years of data were collected, only 2012-2013 is shown here due to space limitations.

A Pearson *r* analysis was computed to assess the relationship between the CELDT listening subscale scores and the ADEPT receptive scores for each school year. The same procedure was conducted for the CELDT speaking subscale scores and ADEPT expressive scores for each school year. There correlations are listed by year and by domain in Table 21.

Table 21

CELDT To ADEPT Pearson r Correlations for Each School Year

School Year	R	p	N
2009–10 CELDT listening to ADEPT receptive	-.264	.087	43
2009–10 CELDT speaking to ADEPT expressive	-.149	.341	43
2010–11 CELDT listening to ADEPT receptive	.229	.025	95
2010–11 CELDT speaking to ADEPT expressive	.250	.015	95
2011–12 CELDT listening to ADEPT receptive	<u>.332</u>	.000	282
2011–12 CELDT speaking to ADEPT expressive	<u>.379</u>	.000	282
2012–13 CELDT listening to ADEPT receptive	<u>.350</u>	.000	348
2012–13 CELDT speaking to ADEPT expressive	**.498**	.000	348

Note. Values that are underlined indicate moderate correlations. Bolded value indicates a strong correlation.

To summarize, for 2009–2010, results showed negative weak correlations. The 2010–2011 correlations were close to moderate between the two variables. For 2011–2012 and 2012–2013, results showed positive moderate to strong correlations between the two variables of listening/receptive and speaking/expressive. Overall, correlations increased and remained positive as the sample size increased, and with each subsequent school year.

RQ3. How well does the student's ADEPT score predict the subsequent CELDT?

Predicative validity is the degree to which a test can predict how well an individual will do in a future situation (Gay et al., 2009). To determine the predictive validity of ADEPT, a Pearson *r* analysis was computed to assess whether ADEPT overall scores in listening and speaking were predictive of the following year's CELDT scores. Correlations are listed in Table 22.

Table 22

Predictive Validity Pearson R Correlations, 3 School Years

ADEPT To CELDT	r	p	n
2009–10 to 2010–11			
ADEPT receptive to CELDT listening	-.114	.273	95
ADEPT expressive to CELDT speaking	.021	.837	95
2010–11 to 2011–12			
ADEPT receptive to CELDT listening	.185	.002	282
ADEPT expressive to CELDT speaking	.135	.023	282
2011–12 to 2012–13			
ADEPT receptive to CELDT listening	-.123	.022	348
ADEPT expressive to CELDT speaking	.077	.152	348

Note. Values that are underlined indicate negative correlations.

The Pearson r Correlations for the receptive/listening and expressive/speaking domains were negative to weak between the two tests for all 3 school years. Receptive/listening results ranged from a low of -.114 to a high of .185. Expressive/speaking results ranged from a low of .021 to a high of .135. Overall predictive validity was weak for all 3 school years.

In sum, ADEPT was found to be reliable at the first two levels of the receptive and expressive domains with alpha scores above .84. The third level had optimal reliability for the expressive domain .82 but less than adequate for the receptive domain .53. Concurrent validity in the receptive/listening and expressive/speaking domains was negative to weak for 2009-10 and 2010–2011 school years. In 2011–2012 and 2012–2013, moderate to strong correlations were found. Predictive validity was weak for all school years, from 2009–2013. An intercorrelations chart showing CELDT and ADEPT subscale scores is available in Appendix B.

CHAPTER 5

DISCUSSION AND IMPLICATIONS

Introduction

The literature pertaining to testing indicates that all tests must be valid and reliable for the purposes intended (AERA et al., 1999). These two properties are the foundation for test creation, test use, and decision making with test results.

Language is a complex construct to assess, yet many English Language Proficiency (ELP) tests have been developed pre and post No Child Left Behind (NCLB). Prior to NCLB, ELP tests were designed without any alignment to standards, curriculum, and textbooks. Once the California English language development standards were put in place, the ELP tests had to be revised or updated to be in alignment for accountability purposes. The California English Language Development Test (CELDT) was created to provide summative information about English language learner (ELL) progress on an annual basis and is a mandated requirement. Formative assessment of ELL progress is not a mandate, but is critical to teachers because it can inform their instruction during the year before the annual CELDT testing.

This study was aimed at validating a formative assessment called A Developmental English Proficiency Test (ADEPT) by first determining the reliability of the test and then comparing it to the California English Language Development Test (CELDT) to establish concurrent and predictive validity.

This chapter provides a discussion of the findings in relation to the research literature and theoretical framework. Additionally, the limitations of the study are discussed and recommendations for future studies are presented.

This study proposed to measure the reliability and validity of a classroom test of listening and speaking skills in English. The research questions explored were: (a) What is the reliability of ADEPT? (b) What is the overall concurrent validity of ADEPT and CELDT in the listening and speaking subscale scores? and (c) How well does the students ADEPT score predict the subsequent CELDT?

Key findings showed ADEPT to have good reliability in levels 1 and 2 in the receptive and expressive domains. In the level 3, reliability was poor for the receptive domain, but the expressive domain was good. The fourth level was not analyzed due to the small sample size.

Concurrent validity was established as moderate to strong depending on the school year, but predictive validity was negative to weak for all 3 school years. These findings correspond with the findings of the ELP tests reviewed in Chapter Two. There are strengths and weaknesses in reliability and validity with all ELP tests, which makes it imperative for educators to know how the results are being used in decision making for students. A difference with the test for this study is that ADEPT is to be formative rather than summative. ADEPT was not designed for accountability purposes.

Discussion of Findings

This study found ADEPT to be a reliable measure of listening and speaking in English in two levels and in the expressive domain for level 3. Results showed internal consistency correlation coefficients that ranged from a low of .53 to a high .97, and a median of .86. The lowest correlation coefficient was in receptive level 3, which had two items dropped due to mean levels of 1.00. Level 4 was not analyzed for reliability due to the small sample size. The reliability results of ADEPT are similar to the Basic Inventory of Natural (BINL), found by

Vecchio and Guerrero (1995), with high reliability correlations up to .925. ADEPT analyses also yielded results similar to those of the LAS-Oral, which had high overall correlation coefficients ranging from .87 to .89, but the listening subtests had a much lower range of correlation coefficients, .48 to .38, which is similar to the intermediate receptive level of ADEPT, with a correlation coefficient of .53. The BINL and LAS-Oral were developed previous to NCLB and the CELDT, so they were not aligned to the California English language development standards, but ADEPT was aligned to CELDT.

Although these results indicate some possible challenges to creating a reliable test, the item statistics for each level of ADEPT provide valuable information for rewriting test items. For example, in level 1, receptive item 13 had a low mean of .38, $n = 113$. The item reads, "Touch your elbow." The correct response is observing the student as he/she touches his/her elbow. It is difficult to know when a student understands the names of body parts and can respond nonverbally by pointing to that part. This particular item needs to be analyzed for content validity through an alignment study or perhaps a sensitivity review.

Similarly, in receptive level 3, items 6 and 10 were dropped due to mean levels of 1.00. For item 6, while looking at a picture, the student is prompted to point to the person in the picture who is asking a question. Item 10 requires looking at a picture, and then pointing to the person who has *not* done something. Both of these items require that the student understand the gestures associated with these behaviors, because there is no language prompt to indicate that a question is being asked or that something has not been done. This example would be considered the type of language that Cummins (1981) calls Basic Interpersonal Communication Skills (BICS), because it is highly contextualized—meaning the context or, in this case, the picture

provides clues to the meaning. The results seem to indicate that these two items are just too simple because every student responded correctly. A careful review of these items would include a sensitivity review or criterion-related study to identify if there is an issue with content validity and/or internal structure. There could be too many context clues in the picture as well.

To summarize, the ADEPT test is reliable in levels 1–2 in the receptive and expressive domains and in level 3 expressive, but a larger sample is needed to determine reliability for level 4.

The concurrent validity results for school years 2009–2010 and 2010–2011 indicated negative to weak correlations for the listening and speaking domains, and the sample sizes were smaller than in subsequent school years. The correlations for 2011–2012 and 2012–2013 were moderate to strong. The range of correlations over the three years was a low of -.264 to a high of .498. Correlations were higher with larger sample sizes and with the most recent school years, which reflects the increasing population of English language learners (ELLs) in the district.

Correlations for predictive validity were negative to weak, with the range being -.114 to .185, indicating no predictive validity between the two tests. It is possible that the time between the tests are too far apart with ADEPT scores from November and the CELDT scores July to October of the following year. The Spring (March) ADEPT scores were collected but not used in the analysis due to administration protocol that states only missed items are tested in the second testing window not the full battery of items. This protocol could constitute an internal threat with the instrumentation of ADEPT. However, because the test is designed as a classroom formative assessment, only testing missed items in the second administration makes sense instructionally. Teachers want to know if their students are making progress, so they would not

want to reassess items that students already got correct in the first administration of the test. This protocol assumes that once students learn a grammatical form they do not forget it so there is no need to reassess previously correct items.

An additional threat to internal validity was found when the data file was organized for the validity analyses. An assumption was that the administrators of the test followed the protocol recommended in the ADEPT assessment manual, which was not found to be the case. Thus, the district established a different protocol for determining the level of ADEPT to begin the test. In some cases, the administrator clearly followed the ADEPT protocol; in other cases, it seemed that the district protocol was followed. In addition, although some students reached the benchmark in a particular level, the subsequent level was not tested. When contacted regarding this inconsistency, the district suggested that the teachers or test administrators ran out of testing time. Another suggested reason was that the district's policy states that students can be considered for redesignation when they reach the intermediate level of proficiency so teachers may not see the value in continuing the test after level 3 intermediate. When the ADEPT protocol is followed, student's scores should continue to increase, as only missed items are assessed in the second administration. Also, from one school year to the next, a student's proficiency level should not decrease. The student's proficiency level should be the same as the previous year or higher. The data showed that if the ADEPT protocol had been followed strictly, several students would have reached level 4.

All tests have limitations, particularly English Language Proficiency (ELP) tests, due to varying definitions of language proficiency, which create uncertainty for the creation of language test interpretation and use of test scores (Stokes-Guinan & Goldenberg, 2011). Many of these

tests are based on Item Response Theory, which assumes that examinees have a latent trait; but, in this case, there is no clear understanding of that trait because the construct of language is so complex. ELP tests are based on different theories requiring different formats so the entire scope of language cannot be assessed in one test. Each test has a different scale for determining proficiency in English. As several researchers have suggested, multiple measures must be used to fully understand a student's proficiency with the English language. Because ADEPT is reliable in the first two levels and at level 3 expressive, and is overall moderately concurrent with the CELDT, the researcher believes it can be recommended for use as a formative assessment and as a multiple measure.

This study makes a contribution to the field of English language development testing because there are fewer classroom assessments than standardized tests of ELP. The literature review revealed only three classroom assessments of ELP with only one based on the California English Language Development Standards. Standards are only one example of student outcomes that can then be measured; but, as sociocultural theory suggests some aspects of language cannot be measured by traditional test forms. For example, is it possible to create a test that assesses social interaction? The researcher found one test, Maculaitus, which claims to measure the contextualized use of language within specific types of situations, but whether that constitutes a measure of social interaction is unknown. According to the National Research Council (2011), in the 2009–2010 school year, this test was not among the eight tests used in 40 states of the United States, so the popularity of its use is unknown. ADEPT has been used widely in California, so many other districts—if willing—could provide the necessary data to replicate this study.

The study is significant because the purpose of providing teachers with ADEPT is to give a tool to inform instruction. Because the concurrent validity did show a moderate to strong relationship, teachers can feel confident that some similar skills are being measured between ADEPT and CELDT in the listening and speaking domains. Although predictive validity was not established, a pattern was discovered over the successive years of the test. Results showed that if ADEPT scores increased, the following year's CELDT scores increased, which may be a result of instruction. Highland District also reported that when ADEPT was used with the results of the district language arts benchmark assessments, ADEPT Level 3 expressive was predictive of advancing a level on CELDT in the following year. This finding by the district has not been proven statistically but could provide the basis for another study.

Although this area of research is growing, the literature highlights a limitation of using classroom assessments for English language proficiency because often these types of tests require teachers to make judgments about a student's performance. The Structured Oral Language Observation Matrix (SOLOM) requires the teacher to rate or score a student's performance on an oral proficiency rating scale, which presents a challenge due to teachers' interpretations of the scale descriptions. Additionally, with the ELD Classroom Assessment, teachers did not apply the scoring criteria consistently, which impacts the reliability of the test (Llosa, 2008).

Recommendations

Several researchers are recommending more formative assessment as part of an overall system of assessing English Language proficiency (Abedi, 2009; Bailey, 2007; Bailey & Huang, 2011). In contrast to using ELP tests for multiple purposes, classroom assessments may fill the

gap of providing more formative information to teachers and providing a more comprehensive view of students' language proficiency.

The first recommendation is to measure the internal consistency of ADEPT level 4 to determine the reliability and, if results are optimal, to proceed with validating the entire test. Another study would have to be performed with a larger sample size than was obtained in this study.

The second recommendation is to revise some ADEPT items based on the item statistics for reliability that identified problems with certain items. The two items in level 3 receptive with a mean level of 1.00 need to be rewritten—which could possibly require a new picture as well. This effort would require field-testing of rewritten items and then another validity study.

A third recommendation is to replicate this study using other standardized English Language Proficiency (ELP) tests. States use different ELP tests as their accountability measure, thus concurrent and predictive validity might be established with other tests. The National Research Council (2011) has published a review of eight tests used by 40 states in 2009–2010, which are administered to approximately 75% of the English language learner students in the country. If validity is found with other states' ELP tests, ADEPT could be recommended as part of their assessment systems.

A fourth recommendation is to reorganize and recalculate these results by eliminating any records that clearly indicate that ADEPT protocol was not followed. This effort would entail removing any records in which the ADEPT scores decreased from the Fall to Spring administration in the same year or from the Spring of one year to the Fall of the next year. Scores on ADEPT should always increase, as that is the design of the test.

This study was quantitative in nature and only focused on the validity of ADEPT. Numerous studies of a qualitative nature could be conducted that would answer questions such as:

1. How does ADEPT inform a teacher's instruction?

2. How does ADEPT influence a teacher's knowledge regarding English language development?

3. How does the professional development provided with ADEPT impact a teacher's knowledge about English language development?

4. What is the relationship between teacher judgments on ADEPT and teacher judgments on CELDT?

These qualitative questions remain of interest to the researcher.

Requiring districts to use reliable and valid tests for ELP assessment is important because these criteria are the cornerstones of ethical testing. Classroom assessment of English language proficiency continues to be a growing area of research. Classroom assessments should be a part of a district's assessment system, as they can be used to assess students' mastery of more standards in a standards-based system, and these types of assessments can be conducted in more authentic and meaningful ways in the classroom (Llosa, 2011). The purpose of multiple measures is to get a broader view of students' abilities, and there are limitations to high stakes standardized tests. For this reason, using a classroom assessment such as ADEPT is supported by the literature as a way of gaining a comprehensive picture of what ELL students know and are able to do (Abedi, 2008).

Table A1

Summary of Four additional ELP tests

Test Grades/ages	Purpose	Domains measured	Proficiency levels
Bilingual Syntax Measure BSM I- Grades K–2 BSM II- Grades 3–12	To provide a measure of oral language proficiency	Listening and speaking	Level 1: No English to Level 5: Proficient English. BSM II has the same four levels as BSM I but levels 5 and 6 are labeled Proficient English I and Proficient English II respectively.
Language Assessment Battery K–2	To assess reading, writing, listening comprehension and speaking in English and Spanish	Listening, speaking, reading and writing	N/A
Maculaitis K–12	To measure the contextualized use of language within specific types of situations.	Reading, Writing, Listening & Speaking	Results are converted into five levels of oral language competency and five levels of literacy.
Language Proficiency Test Series LPTS K–12		Listening, Speaking, Reading, Writing,	The results are reported as two levels of oral language proficiency and four levels of literacy.

Note. Adapted from Current Language Proficiency Tests and Their Implications for Preschool English Language Learners by A. Esquinca, D. Yaden and R. Rueda, 2005, *Proceedings of the 4th International Symposium on Bilingualism*, Copyright 2005 by Cascadilla Press.

Table B1

2009-2010 *Intercorrelations among the* CELDT *and* ADEPT *Subscale Scores in Listening and Speaking*
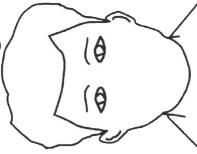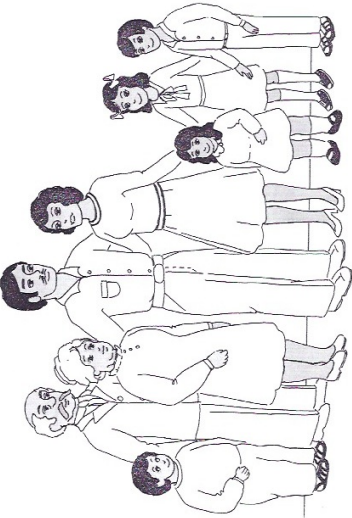
		09–10 CELDT Proficiency Level Listening	09–10 CELDT Proficiency Level Speaking	09–10 ADEPT FALL Overall Receptive	09–10 ADEPT FALL Overall Expressive
09-10 CELDT Proficiency Level Listening	Pearson Correlation	1	.356	-.264	-.264
	Sig. (2-tailed)		.000	.087	.087
	N	43	.43	43	.43
09-10 CELDT Proficiency Level Speaking	Pearson Correlation	.356	1	-.149	-.149
	Sig. (2-tailed)	.019		.341	.341
	N	43	43	43	43

		10–11 CELDT Proficiency Level Listening	10–11 CELDT Proficiency Level Speaking	10–11 ADEPT FALL Overall Receptive	10–11 ADEPT FALL Overall Expressive
10-11 CELDT Proficiency Level Listening	Pearson Correlation	1	.519	.229	.229
	Sig. (2-tailed)		.000	.025	.025
	N	95	95	95	95
10-11 CELDT Proficiency Level Speaking	Pearson Correlation	.519	1	.250	.250
	Sig. (2-tailed)	.000		.015	.015
	N	95	95	95	95
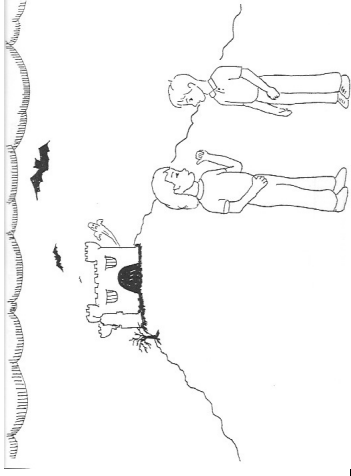
		11–12 CELDT Proficiency Level Listening	11–12 CELDT Proficiency Level Speaking	11–12 ADEPT FALL Overall Receptive	11–12 ADEPT FALL Overall Expressive
11-12 CELDT Proficiency Level Listening	Pearson Correlation	1	.236	.332	.332
	Sig. (2-tailed)		.000	.000	.000
	N	282	282	282	282
11-12 CELDT Proficiency Level Speaking	Pearson Correlation	.236	1	.379	.379
	Sig. (2-tailed)	.000		.000	.000
	N	282	95	282	282

Appendix C

Table C1

Sample Dichotomous Constructed Response Test Items from CELDT and ADEPT

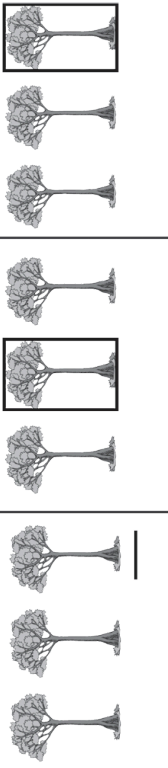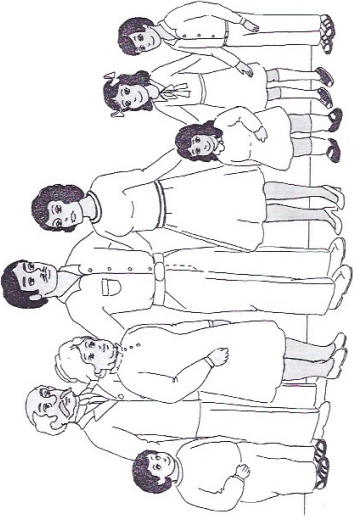| CELDT	ADEPT
Kindergarten Through Grade Two — Listening **1) Following Oral Directions** **ELD Standard:** Respond to simple directions and questions by using physical actions and other means of nonverbal communication (e.g., matching objects, pointing to an answer, drawing pictures). Draw a nose on the boy's face. [Correct response: A nose on the face]	**Beginning Receptive Items** **Family, clothing, and color vocabulary** **Teacher Prompts** 1.24 Show me the father. 1.25 Point to the little brother. 1.26 Get the crayons. Color his sweatshirt green. Acceptable Student Responses 1.24 Indicates the father. 1.25 Points to the smallest boy. 1.26 Colors sweatshirt green.

Kindergarten Through Grade Two — Speaking
3) Oral Vocabulary
ELD Standard:
Begin to speak a few words or sentences by using some English phonemes and rudimentary English grammatical forms (e.g., single words or phrases).
Say: *What is this?*

[Possible answers: Backpack, school bag]

Beginning Expressive Items
Survival Language
Teacher Prompts
1.1 What is your first name?
1.2 What is your last name?
1.3 How old are you?
1.4 What is your teacher's name?
1.5 What is this? (*Point to the table*)
1.6 What do we do with this? (*Show a pencil or pen.*)

Acceptable Student Responses
1.1 Responds with first name.
1.2 Responds with last name(s).
1.3 States age (could be one word only).
1.4 Responds with teacher's name.
1.5 A table; Table; It's a table.
1.6 Write; Draw; We write; We draw.

Kindergarten Through Grade Two — Speaking
4) Speech Function
ELD Standard:
Actively participate in social conversations with peers and adults on familiar topics by asking and answering questions and soliciting information.

Say *Now I am going to tell you about some situations that could happen to you. Then, tell me what you would say.*
Say *You are going to ask a classmate to read with you. What would you say?*
[**The function is making a request.** The student might say, "Will you please read with me?" or "I need a partner. Would you please read with me?"]

Intermediate Expressive Items
Question formation: Present with "does" or "do"
Teacher Prompt
3.9 The girl wants to know how much it costs to go into the haunted house.

What does she ask the other child?

Acceptable Student Responses
3.9 How much does it cost (to go into the haunted house)? OR Do you know how much it costs (to go into the haunted house)?

Beginning Expressive Items
Survival Language
Teacher Prompts
1.7 Count the sisters. How many sisters are in the family?
1.9 How many people are in the family altogether?

Acceptable Student Responses
1.7 Two
1.9 Eight

Beginning Expressive Items
Survival Language
Teacher Prompts
1.8 Who is this? (*Point to the mother.*)

Grades Three Through Five — Listening
6) Following oral directions
ELD Standard: Begin to speak with a few words or sentences, the using some English phonemes and rudimentary English grammatical forms (e.g., single words or phrases). **Scoring:** This question was scored as Incorrect or Correct.
Say: Choose the picture that shows a box around the last tree. Mark your answer. Pause.

Grades Three Through Five — Speaking
7) Oral Vocabulary
ELD Standard:
Begin to speak a few words or sentences by using some English phonemes and rudimentary English grammatical forms (e.g., single words or phrases).

101

Scoring:

This question was scored as No Response, Incorrect, or Correct.

Say: What is this?

[Correct answer: Pear]

Grades Three Through Five — Speaking

8) Speech Functions

ELD Standard: Orally communicate basic needs (e.g., "May I get a drink of water?"). **Scoring:** This question was scored using the "Speech Functions" rubric (0–2) found in Appendix A. Sample student responses are provided below.

Say: *You are drawing a picture. You want to borrow a blue marker from your friend. What would you say to your friend?* [The function is making a request. The student might say, "Can I borrow your marker?" or "Is it OK if I use your marker?"]

Acceptable Student Responses
1.8 The mother, Mother, Mom

Early Intermediate Expressive Items
Routine question
2.4 This girl really needs to go to the bathroom. What does she ask the teacher?
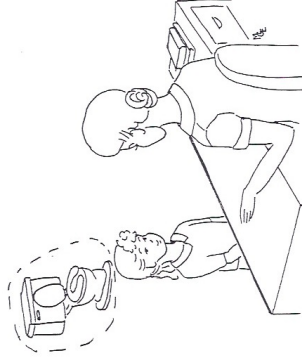
Acceptable Student Responses
2.4 May I go to the bathroom? OR Can I go to the bathroom?

Grades Six Through Eight — Listening
11) Following oral directions
ELD Standard: Restate and execute multiple-step oral directions.
Scoring: This question was scored as Incorrect or Correct.
Say: Choose the picture that shows a line connecting the bus to the truck. Mark your answer.
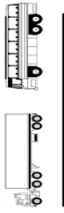
Beginning Receptive Items
Food vocabulary and common verbs

Teacher Prompts
I am going to ask you to pretend to do things. Use the picture to show me how you would do these things.
1.36 Pick up the pie, open the oven, and put the pie into the oven.
1.37 Get the milk. Pour the milk in the glass.
1.28 Drink the milk.
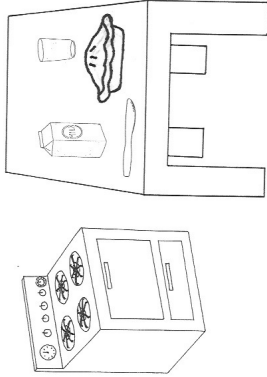
Acceptable Student Responses
1.36 Pretends to pick up the pie, open the oven, and put the pie into the oven.
1.37 Gets the milk. Pours the milk in the glass.
1.38 Drinks the milk.

103

Appendix D

Table D1

Item Statistics Level 1 Receptive

	Mean	Std. Deviation	N
r1	.86	.350	113
r2	.83	.376	113
r3	.82	.383	113
r4	.84	.368	113
r5	.77	.423	113
r6	.77	.423	113
r7	.75	.434	113
r8	.77	.423	113
r9	.81	.398	113
r10	.74	.439	113
r11	.80	.404	113
r12	.80	.404	113
r13	.38	.488	113
r14	.83	.376	113
r15	.73	.448	113
r16	.81	.391	113
r17	.65	.480	113
r18	.85	.359	113
r19	.84	.368	113
r20	.72	.453	113
r21	.72	.453	113
r22	.75	.434	113
r23	.79	.411	113
r24	.83	.376	113
r25	.76	.428	113
r26	.73	.444	113
r27	.64	.483	113
r28	.81	.391	113
r29	.70	.461	113
r30	.51	.502	113
r31	.72	.453	113
r32	.85	.359	113
r33	.66	.475	113
r34	.73	.448	113
r35	.72	.453	113
r36	.80	.404	113

r37	.81	.391	113
r38	.80	.404	113
r39	.89	.309	113
r40	.80	.404	113
r41	.87	.341	113
r42	.86	.350	113
r43	.85	.359	113
r44	.77	.423	113
r45	.92	.272	113
r46	.81	.391	113
r47	.63	.485	113
r48	.72	.453	113

Table D2

Item Statistics Level 1 Expressive

	Mean	Std. Deviation	N
e1	.88	.320	113
e2	.81	.398	113
e3	.81	.398	113
e4	.89	.309	113
e5	.80	.404	113
e6	.77	.423	113
e7	.72	.453	113
e8	.87	.341	113
e9	.82	.383	113
e10	.65	.480	113

Table D3

Item Statistics Level 2 Receptive

	Mean	Std. Deviation	N
r1	.89	.318	300
r2	.98	.128	300
r3	.98	.151	300
r4	.98	.140	300
r5	.97	.171	300
r6	.96	.188	300
r7	.86	.348	300
r8	.80	.398	300
r9	.88	.326	300
r10	.60	.491	300
r11	.96	.204	300
r12	.88	.322	300
r13	.97	.180	300
r14	.92	.272	300
r15	.92	.272	300
r16	.97	.161	300
r17	.80	.403	300
r18	.73	.445	300
r19	.84	.364	300
r20	.90	.301	300
r21	.80	.401	300
r22	.77	.419	300
r23	.87	.333	300

Table D4

Item Statistics Level 2 Expressive

	Mean	Std. Deviation	N
e1	.89	.62	300
e2	.98	.69	300
e3	.98	.66	300
e4	.98	.54	300
e5	.97	.47	300
e6	.96	.47	300
e7	.86	.70	300
e8	.80	.74	300
e9	.88	.55	300
e10	.60	.45	300
e11	.96	.46	300
e12	.88	.45	300
e13	.97	.50	300

Table D5

Item Statistics Level 3 Receptive

	Mean	Std. Deviation	N
r1	.97	.170	68
r2	.81	.396	68
r3	.84	.371	68
r4	.62	.490	68
r5	.97	.170	68
r7	.99	.121	68
r8	.96	.207	68
r9	.97	.170	68
r11	.87	.341	68
r12	.60	.493	68
r13	.68	.471	68
r14	.94	.237	68
r15	.88	.325	68
r16	.96	.207	68
r17	.53	.503	68
r18	.54	.502	68

Note. Item No. 6 and No. 10 were dropped due to the mean score of 1.00.

Table D6

Item Statistics Level 3 Expressive

	Mean	Std. Deviation	N
e1	.50	.504	68
e2	.25	.436	68
e3	.66	.477	68
e4	.49	.503	68
e5	.49	.503	68
e6	.49	.503	68
e7	.38	.490	68
e8	.41	.496	68
e9	.25	.436	68
e10	.26	.444	68
e11	.54	.502	68

REFERENCES

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, *8*(3), 231–257. doi:10.1207/S15326977EA0803_02

Abedi, J. (2008). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment*, *13*(2-3), 193–214. doi:10.1080/10627190802394404

Albers, C. A., Kenyon, D. M., & Boals, T. J. (2008). Measures for determining English language proficiency and the resulting implications for instructional provision and intervention. *Assessment for Effective Intervention*, *34*(2), 74–85.

Alderson, J. C., Krahnke, K. J., & Stansfield, C. W. (1987). *Reviews of English language proficiency tests.* Washington, DC: Teachers of English to Speakers of Other Languages.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Evaluation assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*, 114-129.

Bachman, L. F. (2009). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. New York, NY: Oxford University Press.

Bailey, A. L. (2007). Introduction: Teaching and assessing students learning. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 1–26). Los Angeles: Yale University Press.

Bailey, A. L. (2010). Assessment in schools – Oracy. In P. Peterson (Ed.), *International encyclopedia of education* (3rd ed., pp. 285–292). Amsterdam, The Netherlands: Elsevier.

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, *28*(3), 343–365.

Baker, C. & Prys Jones, S. (1998) *Bilingualism and second language acquisition*. (pp. 635-664) In: Encyclopedia of Bilingualism and Bilingual Education, Clevedon, UK: Multilingual Matters.

Baker, C. (2006). *Foundations of bilingual education and bilingualism*. Tonawanda, NY: Multilingual Matters.

Baker, F. (2001). *The basics of item response theory* (pp. 1–176). In: ERIC Clearinghouse on Assessment and Evaluation, 2nd ed. University of Maryland, College Park, Maryland. Editors Carol Boston, Lawrence Rudner

Borkowski, J. W., & Sneed, M. (2006) Will NCLB improve or harm public education? *Harvard Educationl Review, 76* (4), 503-727.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341–366. doi:10.1177/0265532209104666

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

Burger, D., Mauricio, R., & Ryan, J. (2007). *English language proficiency assessment in the Pacific Region. English* (pp. 1–20). (Issues & Answers Report, REL 2007–No. 014). Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Pacific. Retrieved from http:// ies.ed.gov/ncee/edlabs.

Cadiero-Kaplan, K., & Rodriguez, J. (2008). The preparation of highly qualified teachers for English language learners: Educational responsiveness for unmet needs. *Equity & Excellence in Education*, *41*(3), 372–387.

California Department of Education. (2002). *English-language development standards for California public schools: Kindergarten through grade twelve*. Sacramento, CA: California Department of Education, Retrieved from www.cde.ca.gov/be/st/ss/ documents/englangdevstnd.pdf

California Department of Education. (2011). *California English language development test technical report 2009-2010*. Sacramento, CA: California Department of Education, Retrieved from www.cde.ca.gov/ta/tg/el/documents/formftechreport.pdf

California Department of Education. (2011). *California English language development test technical report 2010–11*. Sacramento, CA: California Department of Education, Retrieved from www.cde.ca.gov/ta/tg/el/documents/celdttechreport10-11.pdf

California Department of Education. (2012). *California English language development test information guide 2012-13*. Sacramento, CA: California Department of Education, Retrieved from http://www.cde.ca.gov/ta/tg/el/documents/celdtinfoguide1213.pdf

California Reading and Literature Project. (2006). *ADEPT-A Developmental English Proficiency Test*. San Diego, CA: California Reading and Literature Project.

Conteh-Morgan, M. (2002). Connecting the dots: Limited English proficiency, second language learning theories, and information literacy instruction. *The Journal of Academic Librarianship*, *28*(4), 191–196.

Cross, R. (2010). Language teaching as sociocultural activity: Rethinking language teacher practice. *The Modern Language Journal*, *94*(3), 434–452.

Cummins, J. (2011). Literacy engagement. *The Reading Teacher*, *65*(2), 142–146.

Esquinca, A., Yaden, D., & Rueda, R. (2005). Current language proficiency tests and their implications for preschool English language learners. In J. Cohen & K. Mc Alister (Eds.), *Proceedings of the 4th International Symposium on Bilingualism*. Somerville, MA: Cascadilla Press.

Franklin, T. G. (2011). Accountability of NCLB, student subgroup count, and their combined impact on our public schools. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI No. 3472521)

Frey, S. (2012). Test scores rise, but achievement gaps persist. *EdSource*. Retrieved from *www.edsource.org/...scores-rise-but-achievement-gaps-persist/19590*

Fulcher, G. (2010). *Practical language testing*. London, England: Hodder Education.

Gándara, P., & Baca, G. (2008). NCLB and California's English language learners: The perfect storm. *Language Policy*, *7*(3), 201–216. doi:10.1007/s10993-008-9097-4

Gándara, P., Maxwell-Jolly, J., & Driscoll, A. (2005). *Listening to teachers of English language learners: A survey of California teachers ' challenges, experiences, and professional development needs*. Santa Cruz, CA: The Center for the Future of Teaching and Learning. Retrieved from www.cftl.org

Gándara, P., & Rumberger, R. (2007). *Resource needs for California's English learners)*. Stanford, CA: Institute for Research on Education Policy & Practice.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational Research: Competencies for analysis and applications*. New Jersey: Pearson Education.

Goldenberg, C. (2008). Teaching English language learners: What research says and does not say. *American Educator*, *32*(2), 8–21.

Goldenberg, C., & Coleman, R. (2010). *Promoting academic achievement among English learners*. Thousand Oaks, CA: Corwin.

Gottlieb, M. (2006). *Assessing English language learners: Bridges from language proficiency to academic achievement* Thousand Oaks, CA: Corwin Press.

Grissom, J. B. (2004). Reclassification of English learners. *Education Policy Analysis Archives. 12*(36). Retrieved from http://epaa.asu.edu/epaa/v12n36/ pages 1-38

Grodsky, E., Warren, J. R., & Felts, E. (2008). Testing and social stratification in American education. *Annual Review of Sociology*, *34*(1), 385–404.

Hakuta, K. (2011). Educating language minority students and affirming their equal rights: Research and practical perspectives. *Educational Researcher*, *40*(4), 163–174.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy Report 2000-1). Stanford, CA: The University of California Linguistic Minority Research Institute.

Hargett, G. R. (1998). *Assessment in ESL & bilingual education: A hot topics paper.* (pp. 1–37). Portland, OR: Northwest Regional Educational Lab

Herman, J. L., Bachman, L. F., & Bailey, A. L. (2008). *Recommendations for assessing English language learners: English language proficiency measures and accommodation uses. Development* (pp. 1–26). Los Angeles, CA:National Center for Research on Evaluation, Standards, and Student Testing

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*(2), 219–230.

Institute of Educational Sciences. (2012). *The nation's report card: Reading 2011* (pp. 1–105). Suggested citation: National Center for Education Statistics (2011). *The Nation's Report Card: Reading 2011* (NCES 2012-457). Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Krashen, S. D., & Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom*. Oxford, England: Pergamon Press

Kline, T. (2005). Classical Test Theory. *In Psychological Testing A Practical Approach to Design and Evaluation* (pp. 91-106) Sage Publications Inc.

Landsberg, B. K. (2004). Elementary and Secondary Education Act of 1965. Major Acts of Congress. Retrieved from http://www.enotes.com/elementary-secondary-education-act-1965-reference//legislation

Lantolf, J. P. (1996) Introducing sociocultural theory. In J. P. Lantolf & S. L. Thorne (Eds.), *Sociocultural theory and second language learning* (pp.1-26) New York, NY: Oxford University Press.

Lantolf, J. P., and Thorne, S. L., (2002) *Sociocultural theory and second language learning.* New York, NY: Oxford University Press.

Lantolf, J. & Thorne, S. (2007). Sociocultural theory and second language learning. In B. Van Patten & J. Williams (Eds.), *Theories in second language acquisition: An introduction.* (pp. 201–224.). Mahwah, NJ: Lawrence Erlbaum Associates.

Lavadenz, M. (2011). From theory to practice for teachers of English learners. *The Catesol Journal*, *22*(1), 18–47.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice, 27(3),* 32–42.

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing*, *28*(3), 367–382.

Mason, E. J. (2007). Measurement issues in high stakes testing. *Journal of Applied School Psychology*, *23*(2), 27–46.

McLeod, S. A. (2007). Vygotsky. *Simply Psychology*. Retrieved from http://www.simplypsychology.org/vygotsky.html

Murphy, A., Bailey, A., & Butler, F. (2006) California English language development standards & assessment: Evaluating linkage & alignment. *Study conducted by CTB/McGraw-Hill for the State of California Department of Education.* Retrieved from www.cde.**ca**.gov/ta/tg/el/documents/**linkagealignstudy**.pdf

National Education Association. (2008). *English language learners face unique challenges.* National Education Association. Washington DC. Retrieved from www.weac.org/Libraries/PDF/**ELL**.sflb.ashx

National Research Council. (2011). *Allocating federal funds for state programs for English language learners.* Washington, DC: The National Academies Press.

NCELA (2007). Glossary of terms related to linguistically and culturally diverse students. Retrieved May 8, 2007, from http://www.ncela.gwu.edu/

Nieto, S. (2004). *Affirming Diversity The Sociopolitical Context of Multicultural Education* New York: Pearson Education, Inc.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Reading, MA: Addison Wesley.

Parrish, T. B., Linquanti, R., & Merickel, A. (2002). *Proposition 227 and instruction of English learners in California: Evaluation update.* Sacramento, CA: American Institutes for Research and West Ed

Parrish, T. B., Linquanti, R., Merickel, A., Quick, H. E., Laird, J., & Esra, P. (2002). *Effects of the implementation of Proposition 227 on the education of English learners, K–12 Year 2 Report*. Sacramento, CA: American Institutes for Research and West Ed

Parrish, T. B., Merickel, A., Pérez, M., Linquanti, R., Socias, M., Spain, A…D. Delancey (2006). *Effects of the implementation of Proposition 227 on the education of English learners, K–12: Findings from a five-year evaluation*. Sacramento, CA: American Institutes for Research and West Ed. Retrieved from WestEd website: http://www.wested.org/online_pubs/227Reportb.pdf

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

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373–391.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, *13*(2-3), 108–131.

Solorzano, R. W. (2008). High stakes testing: issues, implications, and remedies for English language learners. *Review of Educational Research*, *78*(2), 260–329.

Spinelli, C. G. (2008). Overcoming learning difficulties addressing the issue of cultural and linguistic diversity and assessment: Informal evaluation measures for English language learners. *Reading & Writing Quarterly*, *24*, 101–118.

Sticht, T. G., & Armstrong, W. B. (1994). *Adult literacy in the United States*: *A compendium of quantitative data and interpretive comments.* San Diego: San Diego Community College District.

Stiggins, B. Y. R., & Chappuis, J. A. N. (2006). What a difference a word makes: Assessment for learning rather than assessment of learning helps students succeed. *National Staff Development Council*, *27*(I), 10–14.

Stokes-Guinan, K., & Goldenberg, C. (2011). Use with caution: What CELDT results can and cannot tell us. *The Catesol Journal*, *22*(1), 189–203.

Student achievement in California: Equity alert 2010 california standards test results. (2010). *Education* (Vol. 2009). California: The Education Trust-West. www.edtrustwest.com

Tanenbaum, C., & Anderson, L. (2010). *Title III accountability and district improvement efforts: A closer look*. U.S. Department of Education

Title III Accountability Report Information Guide. (2009).Sacramento, CA: California Department of Education. Retrieved from http://www.cde.ca.gov/ta/ac/t3/documents/infoguide0809.pdf

Vecchio, A. Del, & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque, New Mexico: Evaluation Assistance Center-Western Region

Walqui, A. (2006). Scaffolding instruction for English language learners: A conceptual framework. *International Journal of Bilingual Education and Bilingualism*, *9*(2), 159–180.

Wolf, M. K., Kao, J., Herman, J., Bachman, L. F., Bailey, A., Bachman, P. L., ... & Chang, S. M. (2008). Issues in assessing English language learners: English language proficiency measures and accommodation uses. Literature Review (CRESST Report 731). Retrieved from National Center for Research on Evaluation, Standards, and Student Testing (CRESST) website: http://www.cse.ucla.edu/products/reports/r731.pdf.

Wolf, M. K., Kao, J. C., Griffin, N., Herman, J., Bachman, L., Chang, S. M., & Farnsworth, T. (2008). Issues In assessing English language learners: English language proficiency measures and accommodation uses. Practice Review (CRESST Report 732). Retrieved from National Center for Research on Evaluation, Standards, and Student Testing (CRESST) website: http://www.cse.ucla.edu/products/reports/r732.pdf.

Zhu, W., Rink, J., Placek, J. H., Graber, K. C., Fox, C., Fisette, J. L., Dyson, B., et al. (2011). PE metrics: Background, testing theory, and methods. *Measurement in Physical Education and Exercise Science*, *15*(2), 87–99.