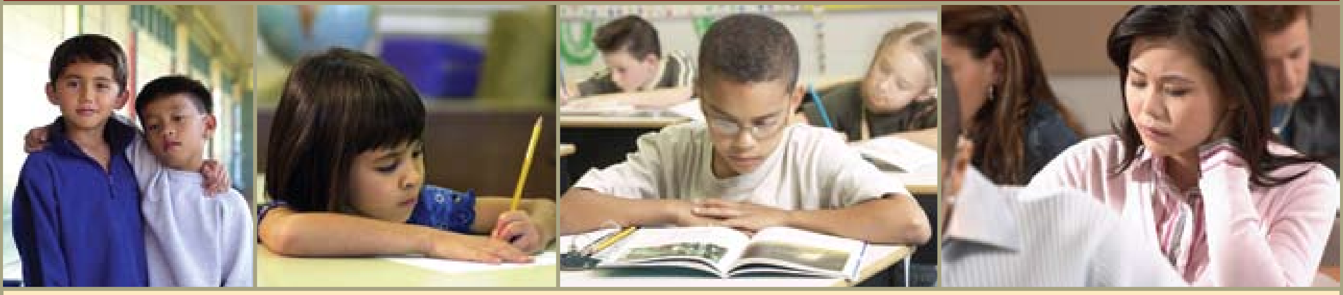
English Language Proficiency Assessment in the Nation:

CURRENT STATUS AND FUTURE PRACTICE



Edited by Jamal Abedi

UCDAVIS
School of Education

UNIVERSITY OF CALIFORNIA, DAVIS • SCHOOL OF EDUCATION

The work reported herein was supported in part by the National Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of the Institute of Education Sciences, or the U.S. Department of Education.

Acknowledgments

This study required the participation of all fifty states, their assessment directors, particularly those involved with the NCLB Title III assessments. Their cooperation and patience is much appreciated, and their dedication to educating young people is greatly admired. We cannot thank the states enough for their collaboration with this project.

Many people generously contributed to the development of this report. We are especially indebted to the chapter authors for their invaluable contributions and for their patience throughout this process. Sue Rigney from the U.S. Department of Education contributed greatly to the quality of this work by providing excellent comments and suggestions. We are so grateful for her support, advice and contribution to this report. Kathleen Leos of the U.S. Department of Education also provided useful comments. We appreciate her support as well. We are also so grateful to Dean Harold Levine for his support, Veena Nambiar, Susan Porter, Jisel Vega, and Lisa Sullivan for their contributions to this work.

Veena Nambiar (a UC Davis doctoral student) helped in this study by reading the initial version of some of the chapters of this report and by providing useful comments and suggestions. Additionally, she collected information on most of the pre-NCLB English language proficiency (ELP) tests and helped with Chapter 7 *(Overview of Existing English Language Proficiency Tests)* of this report. She also contacted many states to inquire about their English Language Proficiency (ELP) assessments policies and procedures and check the accuracy of the data collected for this project related to their states.

Susan Porter (a UC Davis doctoral student) provided assistance to this project by reading some of the chapters and providing useful comments. In addition, she contributed greatly by collecting and summarizing information on pre-NCLB assessments (Chapter 7). She also provided literature review and helped with the definition of the term ELP as opposed to other commonly used terms such as ELD and ESL.

Jisel Vega (a UC Davis doctoral student) contributed greatly to this project by reading the revised chapters, helping with data collection for the existing ELP assessment (pre-NCLB) and by intensive efforts in contacting states and incorporating revisions suggested by the states into the document (Chapter 7). In addition, she contributed greatly in writing Chapter 7.

Lisa Sullivan (a UC Davis doctoral student) contributed to this project by reading different chapters and providing useful comments and suggestions and also by helping with the information regarding existing ELP assessments. She also provided invaluable editorial and formatting comments and suggestions for the entire report.

Mary Courtney (an independent writer) read the entire report and provided invaluable editorial comments and suggestions.

Rita Pope (project coordinator) joined the project at a later phase and provided invaluable assistance in coordinating efforts and proofreading the final version of the report in preparation to print.

We are also grateful to Donna Justice (Director of Communications at the UC Davis School of Education) for her assistance with the processing and printing the report. Her contribution is greatly appreciated.

We also acknowledge the contribution of the following:

Brenda Rinard (a UC Davis doctoral student)

Karine Sarkissyan (a UC Davis student)

Our colleagues from the National Center for
Research on Evaluation, Standards, and
Student Testing at UCLA

Acknowledgments

Authors and Contributors

Jim Bauman, *Center for Applied Linguistics*
Tim Boals, *University of Wisconsin – Madison*
Michael Bunch, *Measurement Incorporated*
Alyssa Buteux, *Educational Testing Service*
Mathina Calliope, *American Institutes for Research*
Yeonsuk Cho, *Ballard Tighe*
Elizabeth Cranley, *University of Wisconsin – Madison*
Steve Ferrara, *American Institutes for Research*
Richard Figueroa, *University of California, Davis*
David J. Francis, *University of Houston*
Cathy George, *California Department of Education*
Margo Gottlieb, *Illinois Resource Center*
Maurice Hauck, *Educational Testing Service*
Kevin Joldersma, *Measurement Incorporated*
Dorry Kenyon, *Center for Applied Linguistics*
Rebecca Kopriva, *C-SAVE*
Julia Lara, *Consultant*
Robert Linquanti, *WestEd*
Ginger Mathews, *Measured Progress*
Christine Mills, *Educational Testing Service*
Veena Nambiar, *University of California, Davis*
Susan G. Porter, *University of California, Davis*
Theodor Rebarber, *AccountabilityWorks*
Mabel O. Rivera, *Texas Institute for Measurement, Evaluation, and Statistics*
Paul Rybinski, *Educational Testing Service*
Robin Scarcella, *University of California, Irvine*
Diana Sewell, *Louisiana Department of Education*
Lisa Sullivan, *University of California, Davis*
Jisel Vega, *University of California, Davis*
Joyce Wang, *Educational Testing Service*
Phoebe Winter, *Consultant*

Table of Contents

Click on topic below to jump to that page

RETURN to TABLE of CONTENTS

Chapter 1

English Language Proficiency Assessment and Accountability under NCLB Title III: An Overview

Jamal Abedi

Understanding the issues concerning instruction, assessment and classification of English language learner (ELL) students is of the utmost importance given the fact that ELL students are the fastest growing student population in the United States. Between 1990 and 1997, the number of United States residents born outside the country increased by 30%, from 19.8 million to 25.8 million (Hakuta & Beatty, 2000). According to a recent report by the U.S. Government Accountability Office, approximately 5 million ELL students were enrolled in schools, representing an estimated 10% of all public school students (GAO, 2006).

The definition of an ELL [or limited English proficient (LEP)] student, as outlined in The No Child Left Behind (NCLB) Act of 2001, (NCLB, 2002) is: (a) age 3 through 21; (b) enrolled or preparing to enroll in an elementary or secondary school; (c) not born in the United States or whose native language is not English; (d) a Native America, Alaskan Native, or a native resident of the outlying areas; (e) from an environment where a language other than English has had a significant impact on an individual's level of English language proficiency; (f) migratory and comes from an environment where English is not the dominant language; and (g) has difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient level of achievement and the ability to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society (NCLB, 2002, Title IX).

The above definition is primarily based on two sources of information: (1) students' language background information and (2) their level of English proficiency. Information on the language background of students (e.g., country of birth, native language, and type and amount of a language other than English spoken at home) comes mainly from the Home Language Survey (HLS). Information on the students' level of English proficiency in speaking, reading, writing, listening and comprehension comes from existing tests of English language proficiency.

Literature on the assessment of ELL students has raised concerns over the validity of information from these sources (see, for example, Abedi, in press). The goal of this report is to present a national view of the status of English language proficiency assessments since, as it will be elaborated later in this chapter, the results of these assessments

RETURN to TABLE of CONTENTS

play a vital role in ELL students' academic careers in many ways including their classification, assessment of their content knowledge, curriculum planning and graduation. We start first with the definition of the concept of *English language proficiency* (ELP).

English language proficiency (ELP) standards, English language development (ELD) standards, and English as a second language (ESL) standards are terms that have been used, often interchangeably, to describe state- or expert-adopted standards that guide the instruction of English learners towards the achievement of English language proficiency. Each of these terms, however, came into use during different time periods, and each was originally based upon constructs that reflected contemporary policies and theoretical frameworks for English learner education.

English as a second language (ESL) is an umbrella term used to describe any one of a number of instructional approaches designed to help English learners acquire English fluency. Most commonly in use during much of the 1990s, ESL was used to describe alternative or supplemental models to bilingual education. Oftentimes, ESL was used to describe sheltered or pull-out English instruction for learners with limited English proficiency (Garcia, 2005). An example of a set of ESL standards developed during this time period is the TESOL ESL Standards (TESOL, 1997).

Since the late 1990s, ELD has more widespread use than ESL. The former denotes "instruction designed specifically for English language learners to develop their listening, speaking, reading and writing skills in English" (NCELA, 2007). In any case, ESL and ELD are used interchangeably, and are both based broadly upon theories of second language acquisition (Wiley & Hartung-Cole, 1998). The California English Language Development (ELD) Standards (CDE, 1999) are examples of standards developed and designed to measure English learners' progress in English language literacy.

While *English language proficiency* has been used for many years to describe benchmarks and levels of English learners' competencies in speaking, writing, listening, and reading, the expression *English language proficiency standards* appears to have become commonplace only since the passage of NCLB and has been adopted by many states that have developed and/or adopted standards specifically to comply with the NCLB Title III provisions. The newest version of the TESOL Standards (2006) reflects the shift towards using this expression, and TESOL adopted this terminology in its recent update of its national standards.

The NCELA glossary (2007) indicates that *ELP* is "often used in conjunction with AMAOs [Annual Measurable Achievement Objectives outlined in NCLB and Title III guidelines]". In spite of these differences, they are used interchangeably in much of the literature on standards-based outcomes and measures for English learners under NCLB. In this collection of reports and discussions on ELP assessment for ELL classification and progress reporting, several of these expressions will be used interchangeably.

Assessing ELLs

The fair and valid assessment of ELL students is among the top priorities on the national educational agenda (Francis, Rivera, Lesaux, Kieffer & Rivera, 2006). To provide a fair assessment for every student in the nation and to assure an equal educational opportunity for all, the NCLB Act mandates reporting of Adequate Yearly Progress (AYP) for all students including four major subgroups, one of which is ELL students. Additionally, NCLB Title III requires states to assess ELL students' level of English language proficiency using reliable and valid measures (NCLB; 2002). Measurement of proficiency is also needed to guide instruction, assessment, classification, placement, progress reporting, and fair decision-making in the accommodation of ELL students.

Assessment impacts ELL students' academic lives in many different ways. In the classroom, assessment of ELL students affects planning of their curriculum and instruction. In particular, ELP assessment plays a major part in the classification and grouping of ELL students. A student's level of English proficiency serves as the most important criteria for the classification that determines their level of proficiency in English and guides the prescription of any needed instruction and instructional materials.

While most states have used commercially-available *off-the-shelf* ELP tests over the years to address this wide array of measurement needs—and some currently use them to fulfill Title III requirements—these assessments' constructs and consequential validity have long been points of discussion (Abedi, in press; De Ávila, 1990; Linquanti, 2001; NRC, 2000; Valdés & Figueroa, 1994, Del Vecchio & Guerrero, 1995; Zehler et al., 1994). For example, reviews of some of the most commonly used language proficiency tests reveal differences in the types of tasks the tests cover and the specific item content of the tests.

The reviews also suggest that these tests are based on a variety of different theoretical emphases prevalent at the

RETURN to TABLE of CONTENTS

time of their development, indicating that the concept of an English language proficiency domain is not operationally defined in many of these tests (see for example, Abedi, in press; Del Vecchio & Guerrero, 1995; Zehler et al., 1994). Furthermore, analyses of data from the administration of some of the existing language proficiency tests reveal problems with their reliability and validity, the adequacy of their scoring directions, and the limited populations on which field-testing samples were based (Abedi, Leon & Mirocha, 2003; Zehler et al., 1994).

There are several major concerns with the conceptual framework of some of the English proficiency tests constructed and used prior to the implementation of NCLB. First and foremost is the divergence in the theoretical frameworks of the tests. These tests are based upon one or more of at least three different schools of thought: (1) the discrete point approach, (2) the integrative or holistic approach, and (3) the pragmatic language testing approach (Del Vecchio & Guerrero, 1995). Consequently the tests provide very different outcome measures. For example, Valdés and Figueroa (1994) observed:

> As might be expected, instruments developed to assess the [English] language proficiency of "bilingual" students borrowed directly from traditions of second and foreign language testing. Rather than integrative and pragmatic, these language instruments tended to resemble discrete-point, paper-and-pencil tests administered orally. (p. 64)

Second, there is a distinction between basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP) (Bailey & Butler, 2003; see also Cummins, 2000). Since ELP tests vary in the extent they gauge academic English, many students could be scoring high in basic interpersonal communication (BICS) without possessing enough cognitive academic language proficiency (CALP) for academic success. After Bailey and Butler (2003) recognized the scope of academic English as "language that stands in contrast to the everyday informal speech that students use outside the classroom environment" (p. 9), they provided its lexical, grammatical, and classroom discourse levels in its three domains, one of which is assessment (2007). In the dual contexts of assessing student achievement and program accountability, it is necessary to determine which tests adequately measure the several types of language proficiency skills needed for success in mainstream English language classrooms (see chapter 2).

With the technical shortcomings in many existing ELP tests, it is no surprise that the NCLB Act upgraded the expectations placed on states regarding assessment of and accountability for the performance of ELL students (NCLB, 2002). Specifically, NCLB Title III requires states to:

1) develop and implement ELP standards suitable for ELL students' learning of English as a second language;

2) implement a single, reliable and valid ELP assessment aligned to ELP standards that annually measures listening, speaking, reading, writing, and comprehension;

3) align these tests with the states' English language development content standards and provide content coverage across three academic topic areas, which include: English/Language Arts; Math, Science, and Technology; and Social Studies as well as one non-academic topic areas related to school environment, such as extra-curricular activities, student health, homework, and classroom management (Fast, Ferrara & Conrad, 2004); and

4) establish Annual Measurable Achievement Objectives (AMAOs) for ELL students that explicitly define, measure, and report on the students' expected progress toward and attainment of ELP goals (see Title 1, Part A § 1111 (b) and Title III, Part A § 3102 (8) and Part A § 3121 (a) (2) and (3)).

These new mandates have generated significant challenges for states with respect to standards and test development; test validity; and accountability policy development and implementation (GAO, 2006; Zehr, 2006; Abedi, 2004; Crawford, 2002).

In response to the NCLB mandate, the U.S. Department of Education provided support to states for developing reliable and valid ELP assessments through the Enhanced Assessment Grant under Section 6112b of the NCLB Act. Four different consortia of states have been developed and are currently implementing ELP assessments that attempt to address Title III assessment requirements. (Reports from these consortia are featured in chapters 3-6 of this report.) The remaining entities are using either their own state-developed tests or some

RETURN to TABLE of CONTENTS

version of commercially available assessments, augmented or off-the-shelf (Zehr, op.cit.). (Chapter 7 provides an overview of tests and includes the states using them.)

Forte (2007) presented a summary of the data that she collected from state Title III assessment between May and October 2006. Based on the results of the survey, of the 33 states responded, 26 states used off-the-shelf ELP tests in 2004-2005 but only seven states continue using these tests in 2006-2007. She indicated that "Other states have adopted consortium-developed tests or created augmented versions of existing instruments to enhance alignment with their ELP standards" (page 15). Forte (2007) also provided very useful information on the currently-used ELP assessments by state. For example, Table 3 in Forte's report shows the ELP assessments used in the past three academic years (2004-2005, 2005-2006, and 2006-2007). We compared our survey responses from states with the data presented in Forte's Table 3 for cross-validation purpose. The information from both sources was generally consistent.

Although federal technical assistance and review of Title III has begun, many state policymakers and education practitioners have voiced the need for enhanced guidance as well as technical assistance and policy support in implementing these assessments and related accountability systems (GAO, op.cit.).

In addition to the four consortia of states, several test publishers have been engaged in major efforts to develop assessments that are consistent with the requirements set forth by the NCLB Title III accountability mandate. These publishers significantly upgraded their existing ELP tests or created new ELP tests that are in line with the NCLB Title III requirements (see chapter 7 for details on tests such as LAS Links and SELP).

Objectives and Content

This report is designed to provide information on the existing ELP assessments and discuss the national efforts in developing new ELP assessments based on the criteria required by NCLB Title III. The report also intends to share research, policy analyses, and technical documents from across the nation which address critical Title III–related issues facing many state policymakers and educational leaders, as well as educational consultants, publishers, and researchers.

To provide the education community with the latest information on the existing and newly developed ELP assessments, state and national experts involved in the development and field testing of the four consortia assessments have provided information on the development of the consortia tests in four of the chapters here. Each chapter includes: 1) the current status of test development and implementation, 2) the alignment of test content with state ELP content standards, and 3) field testing and validation processes. The authors responded to concerns raised by educators and researchers including: 1) differences between the consortia's tests and existing ELP tests, 2) evidence supporting the tests' adequate coverage of content and ELP standards, and 3) research data on the validity of these tests for use in high-stakes assessment and classification of ELLs.

Many existing ELP assessments have been used for many years, and states and districts have information on the content and psychometric properties of these tests, which they may not have for the newly developed ELP assessments. In addition, there may not be enough data to judge the quality of the newly constructed ELP assessments. In response to these concerns, Bauman, Boals, Cranley, Gottlieb and Kenyon (see Chapter 6) compared the existing tests with a newly developed test (ACCESS for ELLs) and also reported major differences between some of the existing tests and a newly developed set of ELP tests (ACCESS for ELLs®, see chapter 6, table 2). Comparisons between the existing and the newly developed English proficiency tests were made in ten areas. Below is a summary of some of their observations that can be applied to other newly developed ELP tests that are based on the NCLB Title III requirements.

1) **Standards-based.** Many of the existing English proficiency tests are not standards-based, whereas the newly developed English proficiency tests are anchored in states' ELP standards.

2) **Secure**. Many of the existing English proficiency tests are non-secure (off-the-shelf tests) and are useful in low-stakes situations. The new generation of ELP assessments are considered secure, high-stakes assessments.

3) **Emphasize academic English**. The social language proficiency focus of many existing ELP assessments is less useful to schools than the new ELP tests that emphasize academic English.

4) **Aligned with academic content standards**. Unlike many of the old generation of ELP tests,

RETURN to TABLE of CONTENTS

the newly constructed ELP assessments are aligned with core academic content standards. However, these new assessments are not tests of academic content, so no content-related knowledge is needed to respond to the newly developed ELP test items.

5) **Includes oral language**. The newly constructed ELP assessments have independent oral language domains consisting of listening and speaking.

6) **Comparable across grades**. Many of the existing ELP assessment batteries offer different tests for each grade cluster, often with no across-grade comparability, whereas many of the newly developed ELP assessments provide opportunities for across grade-level comparisons. Comparability is essential to reporting language proficiency growth, which reflects on student progress, and school accountability.

7) **Tiered within grade levels**. Tests that are available for more than one proficiency level within a grade span can accommodate the variety of skill levels found in ELLs of any age. Some tests are designed to straddle the proficiency level to shorten administration time and lessen examinee frustration.

The newly developed ELP assessments are compliant with the NCLB Title III requirements. It is quite understandable that older ELP assessments may not be compliant with the NCLB requirements since they were developed prior to the implementation of the NCLB legislation. As Bauman, et al. (see Chapter 6) indicated, the prior generation of ELP tests were generally constructed in response to 1970s legislation; therefore, they represent the thinking of behavioral and structural linguistics prevalent at the time.

To better understand the conceptual basis of these tests and how these tests differ, chapter 2, presents principles underlying English language proficiency tests.

Chapters 3 through 6 report the process for development and validation of the four newly constructed ELP batteries created by the four consortia of states. Each of these chapters introduces a consortium, outlines how states' content coverage and ELP standards were used as a base for test-item development, describes the test blueprint, summarizes the process used for test development, and discusses pilot and/or field testing of test items. The authors also present a summary of the process for creating the operational form(s), standard setting, and the validation process. In addition, the chapters outline test administration features, the scoring process, the reporting of results, any data concerning validity of the tests, accommodations used in the assessment, and any technical manuals available.

Specifically, Chapter 3 summarizes efforts by five states, in conjunction with Educational Testing Service and AccountabilityWorks, in developing the Comprehensive English Language Learning Assessment (CELLA). In chapter 4, the development of the Mountain West Assessment by eleven states, in collaboration with Measured Progress, is discussed. Chapter 5 presents the English Language Development Assessment (ELDA) developed by the many members of the State Collaborative on Assessment and Student Standards (CCSSO's LEP-SCASS) and the American Institutes for Research, with assistance from Measurement, Inc. and University of Maryland's Center for the Study of Assessment Validity and Evaluation. Chapter 6 discusses the process of developing ACCESS for ELLs® by the World-Class Instructional Design and Assessment (WIDA), a consortium that grew from three to nine states (plus Washington, D.C.) in partnership with the Center for Applied Linguistics, the University of Wisconsin System, the University of Illinois, and several ELL education and ELP testing experts.

Chapter 7 presents information on the commonly used ELP tests, some of which are still being used by states for Title III reporting purposes. The summaries (which are presented in Appendix A of the report) include the following information: (a) test description, (b) test content, (c) scoring and standard setting (d) alignment to state standards, and (e) any technical/psychometric information (reliability, validity, item-level data) to the extent available.

Beyond ELP test development and implementation, states are also facing complex technical and policy issues in using data from ELP assessments to define AMAO target structures and establish accountability systems under Title III (GAO, 2006). Some of these issues include operationally defining English proficiency, determining reasonable growth expectations, *bridging* the results of different assessments, and setting baselines and annual growth targets for local education agencies (George, Linquanti & Mayer, 2004; Gottlieb & Boals, 2006; Kenyon, 2006; Linquanti, 2004). Chapter 8 discusses methods and research findings on the development and implementation of Title III AMAO policies and systems for California, a state which serves about a third of the nation's ELL population. Specifically,

RETURN to TABLE of CONTENTS

California Department of Education staff and technical consultants from the California Comprehensive Center at WestEd reviewed methods used in 2003 to empirically establish AMAO target structures, then reported on three years of subsequent AMAO data analyses using four years of California ELD Test (CELDT) results for over 1.5 million ELLs. Findings and emerging issues are discussed in chapter 8. In addition, implications for professional development and technical assistance are explored. Lastly, a description is given of what California is doing in response to Title III assessments.

An overall summary and discussion of this report is presented in Chapter 9, along with recommendations for states to make optimal use of their NCLB Title III assessments.

Not Our Purpose

The purpose of this report is to present facts about existing and newly developed assessments. We have no intention of evaluating the quality of existing and newly developed ELP assessments or criticizing any of these tests, whether they were developed prior to NCLB or after the law was implemented. Many of the existing ELP assessments have provided valid assessments for ELL students in the past. These tests were developed based on the information available at the time of test development, and many of them continue to provide very useful information for states, districts, and schools in their assessment of ELL students. We present some evidence on the content and psychometric characteristics of the existing ELP assessments since there were substantial data available for these assessments. For the newly developed tests, however, we have not had as much data on the items to examine and discuss and must wait for data from future test administrations before examining critical characteristics of these tests.

As indicated above, in addition to the four consortia of states, major test publishers also developed ELP tests based on the NCLB Title III requirements. Among them are LAS Link and Stanford English Language Proficiency (SELP) assessments. These tests also look promising. Future research will better judge the quality of the newly developed ELP assessments including those developed by the consortia of states and those prepared by the test publishers.

Most of the materials presented in this report provide information captured at a particular moment in time. ELP tests are rapidly evolving; therefore, information on these assessments quickly changes. Our initial plan was to publish this collection of papers as an edited book. However, since the materials in this collection are extremely time-sensitive we felt that a report format would be a quicker and more efficient way to disseminate information on these tests. This format also gives us the capability of updating the materials more frequently. We welcome feedback and input, which we will incorporate into the subsequent version of this report (both in web and print versions).

While there are many aspects of the new ELP assessments that are very promising, there are several issues that remain unresolved. We elaborate on some of these issues in chapter 9.

We sincerely appreciate the efforts of those who participated in this truly collaborative work. We would like to thank the authors of this report for their generous contributions as well as each state's department of education across the country for their involvement and support. We understand the complexity inherent in the assessments that are used for high stakes testing and accountability purposes and therefore, we value collaboration among the all the departments of education, universities, research groups, test developers, and most of all, the teachers and students who participate in test development. We hope this report helps further the communications and collaborations between states as well as others who are interested and involved in assessments for students, particularly those who are English language learners.

References

Abedi, J. (2004). The No Child Left Behind Act and English-language learners: Assessment and accountability issues. *Educational Researcher*, *33* (1), 4-14.

Abedi, J. (in press). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice.*

Abedi, J., Leon, S., Mirocha, J. (2003). *Impact of students' language background on content-based data: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California: Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

RETURN to TABLE of CONTENTS

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 Education: A design document* (CSE Tech. Rep. No. 611). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Bailey, A. L., & Butler, F. A. (2007). A conceptual framework of academic English language for broad application to education. In A. Bailey (Ed.), *The Language Demands of School: Putting Academic English to the Test*. New Haven, CT: Yale University Press.

Bunch. M. B. (2006). Final Report on ELDA Standard Setting. Durham: Measurement Incorporated.

CDE, (1999). *English language development standards*. Sacramento, CA: California Department of Education.

Crawford, J. (2002). Programs for English Language Learners (Title III). In *ESEA Implementation Guide*. Alexandria, VA: Title I Report Publications.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. *In Schooling and language minority students: A theoretical framework*. Los Angeles: California State University; Evaluation, Dissemination, and Assessment Center.

Cummins, J. (2000). *Language, power, and pedagogy. Bilingual children in the crossfire.* Clevedon, England: Multilingual Matters.

De Ávila, E. (1990). *Assessment of language minority students: Political, technical, practical and moral imperatives.* Paper presented at the First Research Symposium on Limited English Proficient Student Issues, OBEMLA, Washington D.C.

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

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

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

Francis, D. J., Rivera, M. O., Lesaux, N., Kieffer, M., & Rivera H. (2006). *Practical Guidelines for the Education of English Language Learners: Research-based Recommendations for Instruction and Academic Interventions.* Portsmouth: RMC Corporation.

Garcia, E. E. (2005). *Teaching and learning in two languages: Bilingualism and schooling in the United States.* New York: Teachers College Press.

George, C., Linquanti, R. & Mayer, J. (2004). *Using California's English Language Development Test to implement Title III: Challenges faced, lessons learned.* Symposium paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Gottlieb, M., & Boals, T. (2006). *Using ACCESS for ELLs data at the state level: Considerations in reconfiguring cohorts, resetting annual measurable achievement objectives (AMAOs), and redefining exit criteria for language support programs serving English language learners*. (WIDA Technical Report #3). Madison: WIDA Consortium, Wisconsin Center for Educational Research, University of Wisconsin.

Government Accountability Office (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency*. Report GAO-06-815 (July). Washington, DC: Author.

Hakuta, K., and Beatty, A., eds. (2000). *Testing English-language learners in U.S. schools: Report and workshop summary.* National Research Council. Washington, DC: National Academy Press.

Kenyon, D. (2006). *Closing the generation gap: Bridging from the old to ACCESS for ELLs.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Linquanti, R. & George, C. (in preparation). *Implementing Title III AMAOs in California: Findings and lessons learned to foster local accountability & progress*. San Francisco: WestEd.

RETURN to TABLE of CONTENTS

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners.* (Policy Report 2001-1). Santa Barbara: University Of California Linguistic Minority Research Institute.

Linquanti, R. (2004, April, 2004). *Assessing English-language proficiency under Title III: Policy issues and options.* Symposium paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

National Research Council (NRC). (2000). Testing English language learners in U.S. schools. Kenji Hakuta and Alexandra Beatty, Eds. Washington, DC: National Academy Press.

NCELA (2007). *Glossary of terms related to linguistically and culturally diverse students.* Retrieved on May 8, 2007, from National Clearinghouse for English Language Acquisition: http://www.ncela.gwu.edu/expert/glossary.html.

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

Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students. (February 2003). Draft of *Part II: Final Non-Regulatory Guidance on the Title III State Formula Grant Program – Standards, Assessments and Accountability*. U.S. Department of Education.

Rivera, C. (2003). State Assessment Policies for English Language Learners, SY 2000-2001. Paper presented at CCSSO Large-Scale Assessment Conference 2003.

Teachers of English to Speakers of Other Languages, Inc. (1997). *ESL Standards for Pre-K-12 students*. Alexandria, VA: TESOL.

Teachers of English to Speakers of Other Languages, Inc. (2006). *PreK-12 English language proficiency standards.* Alexandria, VA: TESOL.

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

Wiley, T. G. & Hartung-Cole (1998). Model standards for English language development: National trends and a local response. *Education*, *119*(2), pp. 205-221.

Zehler, A., Hopstock, P., Fleischman, H., and Greniuk, C. (1994). *An examination of assessment of limited English proficient students*. Task Order D070 Report. Arlington, VA: Special Issues Analysis Center, Development Associates, Inc.

Zehr, M. (2006). New era for testing English-learners begins. In *Education Week*, v. 25, n. (42, July 12, 2006).

RETURN to TABLE of CONTENTS

Chapter 1

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

Chapter 2

Principles Underlying English Language Proficiency Tests and Academic Accountability for ELLs

David J. Francis and Mabel O. Rivera

Learning is a lifelong process acquired through the interaction of study, instruction, and experience. Language is unique among human capabilities in the roles that it plays in symbolically representing, both internally and externally, our knowledge and experience, goals and aspirations, and feelings and emotions—and in its power to create new knowledge and experiences for ourselves and others. Language is the gateway for learning and the vehicle that facilitates acquisition of new knowledge through direct and indirect interaction with other humans, as well as through the reflective processes of introspection. Language is multi-faceted and varies depending on task demands and content (August & Hakuta, 1997; Bates, Dale, & Thal, 1995; Ellis, 1994). Individuals who become proficient in a language possess a complex set of skills which enable them to effectively express their thoughts and ideas, and to derive meaning from their vast array of personal experiences. Since language is so inextricably linked to our experience and knowledge about the world, it is often difficult to imagine how to separate knowledge from the language used to represent it in memory, to communicate that knowledge to others, and to utilize that knowledge in our daily lives. This isomorphism between language and knowledge is most characteristic of academic forms of knowledge like science, history, and to a lesser extent, mathematics, in contrast to some other forms of knowledge such as athleticism, musical prowess, or artistry. This close coupling of knowledge and language poses unique and difficult challenges for the development of suitable assessments of language ability and content-area knowledge in English language learners (ELLs) in U.S. schools.[1]

With these assumptions about the fundamental role of language in the acquisition and representation of human knowledge in mind, this chapter provides an overview of

[1] For the purpose of this chapter, we use the term English language learners (ELLs) rather than Limited English Proficient (LEP) students. Our intent is to highlight that these students are in the process of acquiring English language skills, as opposed to being limited in their English proficiency. Many of these students may be afforded access to language services due to their language minority status, as discussed in Chapter 1.

RETURN to TABLE of CONTENTS

theoretical frameworks included in the language proficiency literature as it applies to the instruction and assessment of ELL's language. These issues are discussed in light of the No Child Left Behind (NCLB) Act of 2001 and its ramifications for ELLs and the teachers and schools who serve them. We will consider the challenges faced by students learning English in academic contexts and present principles that underlie the use of ELP tests in U.S. schools. After considering the relation between language assessments and content area assessments and how these two systems of assessment might best be integrated to develop a more effective accountability model for ELLs, we will see that an important element to our success is the continued development of English language proficiency tests (ELP tests) and their integration with content-area assessments in an effort to improve instruction and allow ELLs to achieve at higher levels. Through these analyses we intend to show that there are steps that we can take immediately to improve the assessment and accountability systems that we use for monitoring the academic achievement and language proficiency of ELLs and for holding schools, districts, and states accountable for this important and unique subgroup of the student population.

English Language Learners and the Assessment of English Language Proficiency in Educational Settings

Prior to discussing the literature on language proficiency and its implications for the instruction and assessment of ELLs, we present some background on this important and growing subgroup of the U.S. school population. Demographic information on ELLs, information on the process by which ELLs are identified in schools, and the specific sections of the U.S. Education Code that address ELLs have been covered in Chapter 1 of this report (Abedi, 2007) and will not be repeated here. Instead, we focus on the unique challenges that this important subgroup of children faces in achieving success in school. However, we include in this group those students who enter U.S. schools as ELLs, but who, through the course of their experiences inside and outside of school, develop sufficient proficiency to be redesignated as Fluent English Proficient (FEP). They are no longer eligible for language support services under Title III, but represent an important subgroup of students from the standpoint of developing an accountability system that effectively and accurately reports on the performance of ELLs. We will return to the issue of accountability, ELLs, and the role of language and content-area assessments in accountability for ELLs in a later section.

Although we do not wish to repeat the information presented in Chapter 1, there are several elements of the 2001 NCLB law that need to be highlighted for the sake of the present discussion on language proficiency assessment. Congress passed the No Child Left Behind (NCLB) Act with the goal of increasing academic achievement and closing achievement gaps among different student groups, with a particular focus on those who are economically disadvantaged, those who represent major racial and ethnic groups, those who have disabilities, and those with limited English proficiency. Under NCLB, state education agencies are held accountable for the progress of ELLs with regards to both language proficiency and academic content. The Title III section of the law (see Abedi, 2007) supports the need for language instruction, and consequently, requires a fair assessment and evaluation of limited English proficient and immigrant students in oral language, reading, and writing skills in English. An important aspect of the new Title III legislation that could easily be overlooked, but which is a critical element of the new law, is the demand that states align their ELP standards with their academic content standards at each grade. The purpose of this alignment is to ensure that students are developing the academic language that they need in order to succeed in the classroom. This important modification to the law covering the education of ELLs forces states to critically examine the language demands of content-area standards and to ensure that ELLs' language skills are being developed to a level that will enable success in mastering content-area knowledge. This interplay between language and the development and mastery of content-area knowledge is central to any meaningful discussion of instruction, assessment, and accountability for ELLs under NCLB.

ELLs present a unique set of challenges to educators due to the heterogeneity in the population and the central role played by academic language proficiency in the acquisition and assessment of content-area knowledge. Differences among ELLs range widely within the areas of former schooling, first language, socioeconomic status of their families, age, and cultural origin. As a group, ELLs also vary in their academic outcomes. Some thrive in our schools; however, a significant proportion—whether or not formally designated as LEP and thus receiving support

RETURN to TABLE of CONTENTS

services for language development—struggle considerably in developing English proficiency, academic skills, and meeting grade-level standards.

One of the most pressing challenges for educating ELLs is their lack of *academic language* skills necessary for success in school (Scarcella, 2003; Bailey & Butler, 2007). This lack of proficiency in academic language affects ELLs' ability to comprehend and analyze texts in middle and high school, limits their ability to write and express themselves effectively, and can hinder their acquisition of academic content in all academic areas. Given the role that vocabulary and grammar play in academic content areas, ELLs face specific challenges to acquiring content-area knowledge: their academic language, and therefore achievement, lags behind that of their native English-speaking peers (National Center for Education Statistics, 2005). It is important to distinguish academic from conversational language skills, as many ELLs who struggle academically have well-developed conversational English skills. To be successful academically, students need to develop specialized vocabulary that is distinct from *conversational language*.

Academic vs. Conversational Language

Practice and experience in the use of a language set the stage for the development of a repertoire that enables an individual to select appropriate vocabulary in a particular context. We know that, in the case of the English language, proficient individuals are able to discern whether to use conversational or academic language given a particular situation. Conversational language is frequently perceived as the skills and vocabulary an individual retrieves and uses on a daily basis, which becomes natural through practice and experience in a comfortable environment. On the other hand, academic language is regarded as evolving with time and experience, having a direct relationship with the level and quality of instruction that an individual receives. However, one must be cautious in assuming that conversational language is less sophisticated or cognitively demanding than academic language, because both dimensions have different levels of complexity and sophistication.

Solomon and Rhodes (1995) reviewed different perspectives of English academic language and identified two distinct hypotheses that dominated the relatively small body of research literature. The first hypothesis proposed that academic language is a compilation of unique language functions and structures, of which only a few are represented in everyday classrooms. Consequently, these are difficult for ELLs to learn (Valdez, Pierce & O'Malley, 1991, as cited in Solomon & Rhodes, 1995). The second, and most cited hypothesis, distinguishes differences among academic language and conversational language. First proposed by Cummins (1981), the supporters of the second hypothesis argue that conversational language, called Basic Interpersonal Communicative Skills (BICS), is acquired early and is more *context embedded*, making it easier for students to draw on a variety of cues in order to understand the meaning of the language (Cummins, 1981). On the other hand, they proposed that academic language, called Cognitive Academic Language Proficiency (CALP), is *context reduced* and provides only a limited amount of resources from which students can derive meaning.

Further, Scarcella (2003) reviewed the literature on these two hypotheses and proposed a third hypothesis, rejecting the BICS/CALP distinction. Claiming that academic English includes multiple, dynamic, inter-related competencies, Scarcella proposed an alternative perspective of academic English that includes the interaction of phonological, lexical, grammatical, sociolinguistic, and discourse components. Scarcella defined academic English as a variety, i.e., a register, of English that is used in professional books and characterized by linguistic features that are associated with academic disciplines. A *register* is a constellation of linguistic features that are used in specific situational contexts and determined by three variables: field (the subject matter of the discourse), tenor (the participants and their relationships) and mode (the channel of communication, e.g., spoken or written) (Halliday, 1994). According to Scarcella, the register of academic English use includes skills such as reading abstracts, understanding key ideas from lectures, and writing forms such as critiques, summaries, annotated bibliographies, reports, case studies, research projects, and expository essays. Furthermore, Scarcella proposed that academic English includes sub-registers directly related to different disciplines (i.e., science, economics, mathematics) that make academic English impossible to understand with the use of conversational language only.

Most recently, Bailey and Butler (2007) proposed a conceptual framework for the operationalization of the academic English language construct in three language domains: assessment, instruction, and professional development. These authors distinguished academic English

RETURN to TABLE of CONTENTS

from the English used in other settings at three key levels: lexical (including general and specialized lexicons), grammatical (based on its syntactic features), and classroom discourse levels using evidence on these different aspects of the language demands encountered by ELLs in the three language use domains, Bailey and Butler have provided a unique and useful perspective on the different types of English that ELLs are required to master in order to deal successfully with the language demands of school.

Mastery of academic language is one of the most significant ingredients of academic success. Individuals who demonstrate effective use of academic language are able to extract meaning of new content, process it, and add it to previous knowledge. Proficient use of—and control over—academic language in English is the key to content-area learning in our schools. Given the nature of today's academic demands, lack of proficiency in academic language affects students' ability to comprehend and analyze texts, limits their ability to write and express themselves effectively, and can hinder their acquisition of academic content in all academic areas. To be successful academically, students need to develop the specialized language of academic discourse that is distinct from conversational language (Francis, Rivera, Lesaux, & Kieffer, 2006(a); Solomon & Rhodes, 1995). Of course, to understand the acquisition of academic language, we must also understand the complex process of second language acquisition, because for ELLs in U.S. schools, acquiring academic English is an important component of second language acquisition.

Second Language Acquisition

In order to be linguistically and culturally responsive to the needs of the ELL population, teachers must have knowledge of first and second language learning and development, ways of adapting materials, methods of instruction, and assessment. Second language acquisition is the process of learning a language in addition to a native or "first" language. As in many other aspects of learning, the process of becoming proficient in a second language is affected by numerous factors that interact and change constantly in a learner. Estimates of the time required to acquire proficiency in a second language vary considerably, with limited empirical data to inform the debate—beyond descriptions of what is observed given current day practices— and little information on the actual effects of important contextualizing factors such as the age of the student, the level of development/proficiency of the student's first language (L1), the language-learning abilities of the student (e.g., sensitivity to phonological, morphological, and grammatical structures), the approach to and intensity of instruction in L2, and how these various factors may interact with one another. While it is inconceivable that there would not be significant individual differences in the rate of students' acquisition of L2 given the existence of substantial individual differences in the development of L1 proficiency among monolinguals, the range of such individual differences and the factors that moderate them is not currently well-informed by data. (See August & Hakuta, 1997; Cummins, 1981; Hakuta, 2000; Thomas & Collier, 2001 for further discussion and information).

One important aspect of learning a second language is the acquisition of vocabulary. According to Snow and Kim (2007):

> It is estimated that high school graduates need to know 75,000 words in English—that means having learned 10 –12 words every single day between the ages of 2 and 17. ELLs who start even just a few years late need to increase their daily learning rate if they are to match the outcomes of English-only (EO) learners. (p. 124)

Although the rate of vocabulary acquisition in ELLs improves with effective instruction, considerable evidence reflects that outcomes for young ELLs are still significantly behind monolingual English speakers. Optimal learning conditions including effective instruction, exposure to a variety of texts and words, and opportunity to practice, among others, help ELLs to close the gap in vocabulary at a rate necessary to succeed in school. If direct vocabulary instruction in general only ensures exposure to about 300 lexical items per year (Stahl & Fairbanks, 1986), teachers must depend on other resources as well as the self-generative role that vocabulary plays in building students' receptive and productive vocabulary knowledge.

For decades, theorists have tried to explain the process of becoming proficient in a second language in order to understand the instructional needs of ELLs. According to Conteh-Morgan (2002), theories of language acquisition fall within three main categories: *behaviorism*, *innatism*, and *interactionism*. The behavioral theory, proposed by B. F. Skinner, explains language development as being influenced by environmental stimuli where association, reinforcement, and imitation are the primary factors in the acquisition of language. Innatist theories attribute humans

RETURN to TABLE of CONTENTS

the natural ability to process linguistic rules. This school of thought served as the framework for several models of second language acquisition, such as Krashen's Monitor Model (see Krashen, 1988 for a more detailed explanation). This model consists of five hypotheses and attempts to incorporate numerous variables involved in second language acquisition, for example: age, personality, traits, classroom instruction, innate mechanisms of language acquisition, and input, among others. Despite criticism concerning definitional adequacy, Krashen's model had a great impact in the 1980s and motivated research in second language acquisition. Interactionist theories focus on the dynamic relationship between native speakers and language learners (Hymes, 1972) and the interactive nature of language learning. According to this class of language learning theories, learners gain communicative competence through experiences where they learn to correct errors as a result of their exchange of communication with peers. Instructional programs that follow this theory view the teacher as a facilitator for instruction more than the person in control of learning.

English Language Proficiency and ELP Tests

Such theories of second language acquisition have influenced current instruction of second language as teachers and administrators consider the importance of social context, learner characteristics, learning conditions, learning processes, and outcomes during, before, and after instruction. Along with providing leadership for effective instruction, state education agencies have the responsibility to select and/or design appropriate tools to measure the development and acquisition of language proficiency as well as content-area knowledge and skills among their ELL populations. The requirement to assess the acquisition and development of English proficiency among ELLs has changed in statute and, not surprisingly, in operational details over the years since *Lau v Nichols* (1974) first brought these issues to the fore in U.S. public education. At a minimum, the development of an effective assessment system for ELLs will require attention to psychometric principles of test construction, which begins with careful articulation of the purpose and domains of assessment. Under current law governing the education of ELLs (see Abedi, 2007), states must now align their language proficiency standards to their content-area standards and achievement targets. In essence, this change in the education law is an explicit attempt to link the definition of language proficiency under Title III language to the language needed by ELLs to attain proficiency in academic content areas, over and above any level of proficiency required to (a) reach the state's predefined level of proficiency on Title III language assessments, (b) successfully achieve in English language classrooms, and (c) participate fully in society as specified by the definition of Limited English Proficiency under Title IX. This significant change in Title III legislation requires a reexamination of state language proficiency assessments, from the definition of English language proficiency to its assessment, and ultimately to the establishment of standards to define different language proficiency levels.

The term *language* is commonly used in reference to any code comprised of signs, symbols, or gestures that we utilize to communicate ideas and derive meaning. Functionalist theory conceptualizes language as composed of three main components (form, content, and use) and comprised of five skills areas (phonology, morphology, syntax, semantics, and pragmatics) that interact in the effective use of language (Raymond, 2004). According to this theory, the first component of language is *form*, which represents how a language is used (i.e., rules). Within form, an individual learns three skills: phonology, morphology, and syntax. Phonology relates to the sounds included in a language, morphology relates to the roots or units of meaning in a language, and syntax is related to the grammatical arrangement of words in sentences in forms that are acceptable within the language for the conveyance of meaning. Languages differ from one another in all three of these areas. The second major component of language is *content*, which represents the semantics or the meaning of words as they interact with each other. Finally, the third component of language is *use*, which is described as pragmatics, or the relationship between the language and the message in communication.

Individuals who master those five skill areas associated with these three components are able to derive meaning and express their thoughts through the four linguistic domains: *reading*, *listening*, *speaking*, and *writing*. The first two domains are considered *receptive* channels and reflect the ability of the individual to manage different forms of linguistic input from the environment. Note that *reading* and *listening* also include lip-reading or signed language for linguistic input and non-print forms of text, such as braille. That is, we define these input and output channels broadly

RETURN to TABLE of CONTENTS

in considering the many different kinds of language users and language test takers in schools. The domains of *speaking* and *writing* are considered *output* or *expressive* channels that we use to communicate our ideas to others and to react to the stimuli, both internal and external, we experience. Just as in the case of the input domains, the output domains are defined broadly to include signing, symbolic coding, and synthesized speech, such as that used by the famous scientist Stephen Hawking, whose degenerative neuromuscular disorder prevents his voluntary control of the articulatory apparatus. To be considered proficient in a language, an individual must be able to communicate effectively and understand the message conveyed through these different domains. Of course, physical limitations that prevent effective reception or expression of linguistic input through certain channels do not, de facto, limit an individual's capacity for language proficiency. Such characteristics of test takers must be taken into account both at the stage of defining the constructs of interest (e.g., language proficiency) and in designing assessments to best meet the needs of all possible test takers (Bachman & Palmer, 1996).

Language proficiency involves the effective use of language to accomplish different objectives of importance to the language user, and reflects linguistic competencies in multiple dimensions. One dimension of competence is simply the level of accuracy achieved by the language user. In a sense, accuracy concerns the degree to which the language user can successfully communicate information through linguistic output channels, and successfully derive meaning through linguistic input channels. To what extent is the message received equal to the message sent, and vice versa. Accuracy here reflects both linguistic and non-linguistic information that is carried through linguistic channels, e.g., humor, emotion, intent, motivation, all of which are carried in different components of the message, sometimes in the words themselves, sometimes in the timing, intonation, or stress, but all of which make up part of the linguistic input/output of the communication. These non-semantic elements, the broader elements of organization, the intersentential relations, and the cohesion of the message draw on the language user's metacognitive skills that enable use of the different pragmatic skills required to communicate accurately in particular settings.

The purpose of testing English language proficiency is threefold: to determine placement in language programs, to monitor students' progress while in these programs, and to guide decisions regarding when students should be exited from these programs (August & Hakuta, 1997; Kato et al., 2004; National Research Council, 2004). Depending on the results and population's unique needs, ELLs may qualify to receive instruction in several types of programs such as English as a Second Language (i.e., pull-out, one class period, or resource center sessions), Bilingual programs (i.e., transitional, two-way or dual-language programs), or other program models such as Sheltered English or Structured Immersion. ELP tests may also be used during monitoring to document mastery of specific standards in language acquisition. Last, the results of ELP tests may provide information for informed decisions during instruction and determine exit from support programs.

Principles Underlying the Measurement of Language Proficiency

Defining and assessing language proficiency presents numerous challenges to the language and assessment communities and has been the subject of debate among researchers for over two decades. The literature is divided among researchers and theorists who either view language proficiency as the effective interaction of multiple linguistic components or perceive it as one global factor. Other researchers focus on the effective use and control of language as it is affected by the situation in which it takes place (Cummins, 1984; Valdés & Figueroa, 1994).

Bachman and Palmer (1996) provide a helpful distinction for test developers, test users, and test takers alike, in discussing the principles that underlie effective language test development. These authors point out that the goal of language assessment is almost certainly the prediction of the test takers' language use in situations external to the language assessment itself. They use the distinction between testing and non-testing contexts to introduce the notion of target language uses (TLUs) as distinct from language test tasks. TLUs describe the different uses of language that occur in non-testing contexts and that are of interest to test users; whereas test language tasks are the specific kinds of language tasks that test takers encounter in the testing context and that test users will leverage to make predictions about TLUs of interest. Thus, for the language assessment to be useful, the scores obtained by an individual on the basis of interacting with a set of test language tasks must generalize to a set of TLUs of interest reflecting the test takers' language use in contexts other than the language testing context. The idea of test language tasks, which are the language tasks encountered in the testing context, is not to be confused with the notion of Task-

Based Language Assessment (TBLA) (Mislevy, Steinberg, & Almond, 2002). TBLAs of all sorts comprise one form of test language task, but unlike some language test tasks that assess language skills directly in a highly decontextualized testing situation, TBLAs attempt to contextualize the language assessment tasks more to incorporate the sociolinguistic, strategic, and discourse-level competencies that occur in typical language use settings.

At present, it remains an open question which type of assessment provides more accurate inferences about the language proficiency of school-aged ELLs with respect to their abilities to deal with the complex language demands of acquiring content-area knowledge in English. For example, it is conceivable that a TBLA-type assessment would afford more accurate inferences about a student's ability to meaningfully engage in classroom-based discussions about material encountered in texts, instructor-provided lectures, or presentations by other students. It may also predict ability to search out, organize, and synthesize material from different sources in order to develop and communicate a coherent term paper on a topic. In contrast, a more traditional standardized test may afford more accurate inferences about the student's ability to compare and contrast different, non-student-selected texts on a topic; craft a unique position paper on those texts; and answer specific questions about the factual basis and the assumptions that underlie the arguments in those texts. Such a test might better predict the student's performance on the end-of-unit examination, the statewide high-stakes assessment, or a job-related performance assessment where the examinee is constrained in terms of the materials available to solve the problem or the form that the solution can take. It is important that tests be evaluated empirically on the validity of such inferences and not simply on the basis of face-validity arguments.

A number of authors have discussed approaches to developing and evaluating language proficiency tests (see Bachman & Palmer, 1996; McNamara, 2000; 2006; Stoynoff & Chapelle, 2005). The notion of test usefulness has been discussed by many authors, including Bachman and Palmer (1996), who argue that test usefulness is a unifying principle that embodies other important principles of test construction and evaluation. These principles—reliability, construct validity, authenticity, interactiveness, impact, and practicality—represent the primary qualities of tests. For these authors, maximizing test usefulness is a matter of balancing these different, inter-related dimensions. Unfortunately, these dimensions can never be simultaneously maximized, as we hope to elucidate in the discussion below.

Reliability is the essential quality of test scores, without which all other test qualities are irrelevant. Reliability can be thought of as the precision of a test score, reflecting the degree to which a test would yield a consistent score for an individual, whose ability remained unchanged. The easiest way to think about reliability is to consider the hypothetical case of an examinee being given the opportunity to take the test for the first time on either of two separate occasions. Assuming that the examinee's true ability is the same on both occasions, reliability describes the degree to which the examinee would be expected to get the same score on each occasion. The reliability of a score is a theoretical abstraction and cannot be known, but it can be estimated in many ways, such as by looking for consistency in scores across different items or components of the test, having examinees take the test on more than one occasion, and having the examinee tested on alternate forms. Because ELLs often lack specific vocabulary in English which can be reasonably assumed for monolingual students of a given age, differences in the language of alternate test forms which have no effect on the performance of monolingual English-speaking students may alter the performance of ELLs. Similarly, vocabulary which can be assumed familiar to monolinguals may be less well known by ELLs and thus may be accessed with error or inconsistency, such that on one occasion the student can recall the meaning of a critical word in an item and on another occasion is unable to retrieve the meaning of that same word. In such situations, the student might answer the question correctly on the first occasion and be unable to answer it on the second occasion. While many of the factors that affect test reliability for monolingual English-speaking students (e.g., fatigue) will also affect the reliability of test scores for ELLs, these same factors may differentially affect the performance of ELLs because of the unique challenges that working in a second language place on test examinees (e.g., reading in a second language places greater demands on mental effort than reading in a first language, and consequently, the student taking the test in a second language may be more likely to become fatigued and experience the effects of fatigue on performance).

The notion of construct validity has evolved over time, and it is beyond the scope of this chapter to review that history. In current thinking about tests, all forms of validity have been subsumed under the heading of construct validity following the work of Messick (1989). Modern

RETURN to TABLE of CONTENTS

notions about construct validity are built on two fundamental tenets. 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. Thus, it is clear from the foregoing that validity concerns the accuracy of test score interpretations. In the language introduced above from Bachman and Palmer (1996) regarding the distinction between test tasks and TLUs, validity concerns the degree to which the test scores derived from test tasks justify interpretations about the TLUs of interest to us. In other words, to what extent can we justify interpretations about how the examinee will perform in situations of interest outside of the testing context? While reliability concerns accuracy as reflected in the consistency of test scores, validity concerns accuracy of test score interpretations, that is, the veracity of inferences about language behaviors outside the testing context. The fundamental validity question regarding language proficiency tests and ELLs is whether a student who scores in the proficient range of the test can function independently in an English-speaking classroom without specific language supports, just as the fundamental validity question regarding content-area assessments is whether or not a student who meets the passing standard possesses grade-level mastery of the content. Test developers and state assessment specialists have to be concerned that a score on the state math test means the same thing for an ELL as for a monolingual English speaker. If the test carries a significant language load for ELLs but not for monolingual students, then the test measures both language and mathematics knowledge for the ELL, but only mathematics for the monolingual student. In this case, the same score interpretations would not be supported for the two kinds of students.

Authenticity and interactiveness are test qualities that reflect the degree of correspondence between test tasks and the TLUs to which we wish to generalize. Authenticity concerns the degree to which the test tasks are comparable to the ways in which test users will use language outside of the testing context. For example, selecting lists of synonyms and antonyms from lists of possible alternatives is a test task that is low in authenticity because test takers rarely find themselves needing to accomplish such a task outside of the testing context, whereas restating or rewriting a sentence substituting an alternative for a target word or phrase is more authentic because speakers and writers are often confronted with the need to provide an alternative formulation of their communication because of unfamiliarity with a particular word or phrase on the part of the audience. Similarly, the test tasks above could be designed as non-interactive or interactive. If the examinee were told which word or phrase must be replaced, the task would be low in interactiveness. In contrast, a more interactive approach to the same task might be constructed by asking the test taker to determine what needed to be restated based on input from the examiner, such as a question, or comment reflecting a lack of understanding of some, but not all elements. For Bachman and Palmer (1996), interactiveness is gauged by the degree to which the test takers' individual characteristics are involved in successfully completing the test task. The individual characteristics of interest in this instance are the individual's language knowledge, strategic competence, and knowledge of the topic, among other things. Interactive and authentic tasks create a testing situation that more closely resembles the ways in which test takers use language in the real world and thus create a perception among the test taker that the test is relevant to them and will accurately reflect their linguistic competence. If students do not sense that the test is comprised of tasks which are meaningful to them or do not reflect the kinds of ways in which they are called on to use language in school, their performance may be adversely affected due to poor motivation or failure to engage cognitively in the tasks of the test in the same way that they engage cognitively in instruction or other language-based learning activities. Ultimately, such disconnects between the testing situation and real-world uses of language could adversely affect the validity of test scores and could differentially affect the validity of test scores for subgroups of ELLs. For example, one might predict that ELLs with higher levels of language proficiency can engage cognitively in more abstract and decontextualized language tasks, while students with lower levels of language proficiency may not.

The final two dimensions of usefulness discussed by Bachman and Palmer (1996) are impact and practicality. Impact has many dimensions itself and includes impact for the test taker, the test user, as well as larger units, such as the entire educational system. Measuring language proficiency has a direct impact on the test taker in many ways. One important way is in the decisions made regarding the type of support that individuals receive in school.

RETURN to TABLE of CONTENTS

A student's measured level of language proficiency is often the main source of information used in deciding which instructional model will best serve their language and learning needs. Another set of impacts is participation in content-area achievement testing and the assignment of accommodations during instruction and testing in order to obtain a fair picture of a student's content knowledge (Abedi & Hejri, 2004). Often ignored is the impact that test use has on instruction for students as well as the effects of testing on larger educational systems. This process whereby assessment impacts instruction and larger educational systems has been termed washback (Garcia, 2003; Hughes, 1989), which has a decidedly negative connotation, but such impacts are neither inevitable nor are they necessarily negative. For example, it is hoped that the reformulation of Title III to align language standards and language proficiency testing standards with academic content standards will increase the focus on academic language development in instructional settings, which is expected to have positive effects on the educational outcomes of ELLs.

Finally, tests must be practical in order to be useful. This conclusion requires little discussion. Clearly, it is conceivable, given the technology that exists today, to measure students' actual language to very high degrees of precision in naturalistic settings by collecting language samples, transcribing them, and analyzing them through computerized models of language. Indeed, all of a student's written and oral productions and all of their linguistic inputs, both auditory and visual, could be captured and recorded digitally, and synthesized through available technology. Even if such a system were technologically possible, would it be feasible to do so, and would the cost of building such a large, complex, and comprehensive language database for each and every student be manageable? Would such a system lead to a sufficient improvement in precision in measuring students' language so as to offset the significant cost burden of such an assessment system? If the assessment requires too much time, too much money, or is too complex given the other demands on student and teacher time and school budgets, then no matter how *good* the test is, it will not be practical and therefore will not be useful.

Usefulness of tests is not a matter of maximizing each of these dimensions. There are always tradeoffs that must be made in developing tests and selecting them for use in specific contexts for specific purposes. Test developers and test users will need to balance the often competing demands of these different dimensions. Still, some dimensions seem less expendable than others. A test which is unreliable or has poor validity will not be useful regardless of how practical, interactive, or authentic it is. In that sense, the first two dimensions of usefulness take priority over all others. At the same time, no matter how reliable a test may be, if it takes too much time away from instruction or places a prohibitive cost burden on the school or state, it will not be practical and thus cannot be useful in that context. Perhaps with changes in technology or resources, the time and/or cost demands of the test could be reduced so that it became practical. And so it would seem that practicality, like reliability and validity, is a more critical dimension of usefulness in that without it, how a test fares on other dimensions becomes somewhat irrelevant in a context where the test is not practical.

Levels of Proficiency

As in many areas of human cognition, the development of language proficiency describes performance on a continuum, but can also be characterized by levels of proficiency that describe distinct stages of development. These stages or levels of development are somewhat different in second language acquisition than in first language acquisition, if for no other reason than that when individuals acquire a second language as skilled speakers of a first language (L1), they have their knowledge in their first language and their metalinguistic knowledge of their first language on which to draw. Obviously, the degree to which such facilitative effects take place will vary across individuals, based on a variety of factors, but most importantly will vary based on the level of skill acquisition in the first language. For young ELLs, whose first language is still very much in the early stages of development, the ability to leverage their knowledge in L1 will be less than for older ELLs with high degrees of competence in L1 and more substantial world knowledge bases on which to draw.

High levels of language proficiency facilitate the processing and acquisition of new information and allow the individual to derive meaning with less conscious effort. These benefits are apparent both intra- and inter-linguistically. For instance, oral language ability in English is a predictor of reading achievement in English (Snow, Burns, & Griffin, 1998), a fact which is not too surprising given that reading comprehension is essentially a language-based task. Even the decoding aspects of reading are language based, mediated in large part by processes in all alphabetic

RETURN to TABLE of CONTENTS

languages (Ziegler & Goswami, 2005). What is becoming increasingly clear is that proficiency in a first language confers benefits on developing competency in a second language, such as the effect that L1 oral proficiency has in developing oral proficiency in L2 (August & Shanahan, 2006), which may be enhanced when languages share common alphabetic characteristics (Miller et al., 2006), cognates, and morphological structures (Snow & Kim, 2007). Proficiency across two or more different languages may be depicted as a multi-tipped iceberg in which common cross-linguistic proficiency lies beneath the obvious differences of each language (Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 1998). Theories of foreign language learning also point to individual differences in various language and/or cognitive capacities to explain variation in the ability to acquire language (Grigorenko, Sternberg, Ehrman, 2000).

The levels of proficiency commonly identified in ELP assessments focus on stages of language learning that presumably can be mapped to the test takers' instructional needs and language-learning support. While across state tests and state standards there is little consistency in the operational definitions of proficiency levels, it is not uncommon to find levels of proficiency labeled *beginner*, *intermediate*, *advanced*, and *transitional*. Some assessments provide finer gradations in the levels, but in all cases, the highest level is intended to reflect proficiency that is capable of independent participation in mainstream English language classrooms. With the changes to language in Title III regulations calling for the alignment of language standards with the language demands of content-area proficiency, this highest level of proficiency could begin to reflect language proficiency that is necessary to achieve proficiency in content-area knowledge and skill acquisition.

The Role of Language Proficiency in Content-Area Knowledge Acquisition

Having examined some of the history behind language proficiency testing in U.S. public education, theories of second language acquisition, definitions of language proficiency and of academic language, and principles of language test development, it is worthwhile to consider the broader context for language proficiency testing as set forth in *Lau v Nichols*. The overriding purpose behind Title III is to ensure that students with limited English proficiency receive the instruction they need to become proficient in content areas, regardless of their current level of proficiency in English, and that they receive instruction that will lead to the development of English language skills that will enable them to participate fully in U.S. society. While there is much more to the law, it is, in its essence, very simple: develop high levels of English language proficiency among ELLs and provide instruction that allows them to acquire content-area knowledge while they are developing proficiency in English. The NCLB modification that requires states to align their language proficiency standards with their content-area standards highlights that a fundamental objective of English language instruction is to cultivate the language skills needed for proficiency in content-area knowledge.

This linkage of language proficiency standards and testing to content-area standards in testing begs the question: To what extent do state language proficiency assessments currently predict student success on content-area assessments? A second important provision in NCLB exempts ELLs from content-area assessments and from Adequate Yearly Progress (AYP) determinations for one year. This period of exemption is allowed because content-area assessments may not be valid or reliable for ELLs until their English language skills have developed to some, as yet unspecified, level which takes more than one year to reach. In so far as NCLB is predicated on the basis of scientifically informed and data-driven decision making, it is surprising how little empirical work has been done to examine these relationships. In this next section, we will review a recent study that looked at one state's link between language and content assessments and will present some recently completed analyses that examine the role of language proficiency and time in the U.S. in predicting student performance on content-area assessments. These studies will lay the groundwork for the final section of the chapter where we consider how to develop a more useful accountability framework for ELLs.

There are surprisingly few studies that have examined the link between language proficiency assessments and state content-area tests on a statewide or district level. Bailey and Butler (2007) reference some unpublished work in their book on the language demands of statewide accountability tests. A number of studies on test accommodations have also looked at linguistic modifications and the role of language in specific test items. However, most of these studies have looked at performance on NAEP and/or NAEP-like assessments, and have not looked at

RETURN to TABLE of CONTENTS

performance on state accountability assessments (Francis et al., 2006(b)). One report that directly examined links between language and content assessments in one state was released by the National Center on Educational Outcomes in collaboration with the Council of Chief State School Officers and the National Association of State Directors of Special Education (Kato et al., 2004). The report examined data from the state of Minnesota as part of an effort by that state to substitute their ELP test, the Test of Emerging Academic English (TEAE), to stand in for ELLs as their reading/language arts test, instead of the Minnesota Comprehensive Assessment (MCA) which is used in grades 3–5, and the Minnesota Basic Skills Test (BST) which is used in grade 8. The report found that predictions differed between the MCA and BST, with greater predictive power in the TEAE for the younger grade MCA, although the test worked comparably across grades 3–5 and across multiple years of data collection. Kato et al. reported multiple R^2 for grades 3 and 5 that ranged from .54 to .58. While generally strong relationships are indicated in this case, the analysis did not take into account the clustering of students within schools and districts, thus correlations at the student level are likely to be smaller. For example, if 10% of the total variance exists at the school level and another 10% at the district level, then the percentage of variance accounted for at the student level would drop to .34–.38.

It is unclear just how high the predictive relationship should be before substitution of one test for another is justified, whether variance accounted for is the right benchmark on which to base such a decision, and whether such a goal is worthy of pursuit. On the one hand, lowering the testing burden for the student is worthwhile. If some tests are redundant or are not reliable or valid for students with certain characteristics, then these tests should be eliminated or eliminated for students with such characteristics. Of course, if the problem is reliability or validity for certain students, then substitution would not be the right goal. Rather, eliminating the unreliable test with poor validity should be the objective. On the other hand, if the tests were redundant, then we would expect the relationship between tests to be strong. In this case, it would make sense to eliminate one of the two assessments, although it would seem that the more logical test to eliminate would be the language proficiency test, since proficiency on the content-area test is the ultimate goal.

We have begun a similar investigation on the links between language proficiency and content-area assessments using data from a second state, whose name cannot be released at the present time. In this particular instance, we are also investigating the role that years spent in the U.S. (Years in U.S.) plays in predicting the development of language proficiency and proficiency on the content-area tests. The dataset included data from the entire state population of ELLs and former ELLs for two academic years (2004–2005 and 2005–2006).

To investigate the mutually interdependent roles of Years in U.S. and development of competency in English on the development of proficiency in English language arts (ELA) and mathematics (MATH), a series of multi-level regression models was run using PROC MIXED in SAS 9.1 (2006). Models were run separately for ELA and MATH and for each of grades 4–8. Several models were examined. First, *unconditional* models were run to estimate the variability in ELA and MATH performance at the student, school, and district levels. Second, we ran two *conditional* models predicting ELA or MATH from: (1) *Years in U.S.* and (2) *ELP* and *Years in U.S.* together, in order to estimate the additional effects of ELP over and above Years in U.S.. In all models, Years in U.S. was treated as a categorical measure such that separate means were estimated for each level of Years in U.S. in order to allow it to account for the maximum variation attributable to this measure. In modeling the effects of ELP, models were run using the ELP composite scaled score (ELP_SS), which is a single score derived from performance on the four separate domains of reading, writing, speaking, and listening. All models allowed for differences between schools and districts in mean performance levels and took into account the nesting of students within schools, and the nesting of schools within districts. Thus, each model provides estimates of three sources of variability in students' scores, namely, variability due to districts, variability due to schools within districts, and variability due to students within schools. It is the reduction in variability due to these three sources that determines the overall explanatory power of the models.

Figures 1 and 2 are designed to show the relationship between ELP_SS and ELA (Figure 1) or MATH (Figure 2) conditioned on Years in U.S. In Figures 1 and 2, each plotted symbol reflects the performance of an individual student and shows where that student scored on ELP_SS and ELA (Figure 1) or ELP_SS and MATH (Figure 2). Both figures are organized in a similar fashion, such that data for a given grade are presented in a particular column (i.e., grade 4 in the left-most column, up through grade 8 in the

RETURN to TABLE of CONTENTS

right-most column). The rows of each figure correspond to the different number of Years in U.S., with the bottom row indicating one year, the second row indicating two years, etc., up to the top row indicating 5 or more years. Within each row and column is a scatter plot showing the relationship between ELP_SS performance and ELA (Figure 1) or MATH (Figure 2).

In Figures 1 and 2, variability in content-area proficiency increases at higher levels of performance on the ELP assessment as reflected in the greater spread in the cluster of points at the right-hand end of each of the scatter plots in each cell. The relationship does not appear markedly different within a grade for students with different numbers of Years in U.S. (i.e., across rows within a given column). This increase in variability in ELA and MATH associated with increased proficiency in ELP could represent various factors, operating alone or in combination. First, it is possible that students need to reach a particular level of performance on the ELP assessment before ELP performance begins to effect performance on the ELA and MATH assessments. Such a threshold phenomenon could occur if the factors that drive performance on the ELP assessment at the lower end of the scale differ from those that drive performance at the upper levels of the scale, and it is only those factors that drive performance at the upper ends of the scale that are related to ELA and MATH performance. Alternatively, such a threshold effect could also be obtained if the ELA and MATH assessments are not accessible to students at the lowest levels of language proficiency. Of course, the purpose of the ELA and MATH assessments is to cover the respective domain in each grade, not to cover the full range of linguistic competence observed within the ELL population within a grade. Finally, this pattern of results could be obtained if the ELP scale had an insufficient ceiling. In other words, if the range of the ELP assessment could be extended, then the overall relationships would be linear and would not show increased variability at the upper end of the ELP scale.

Another possible interpretation of the pattern of relations shown in Figures 1 and 2 is that of differential validity of the ELA and MATH assessments for students at different levels of performance on the ELP assessment, or if not differential validity, differential utility of the ELA and MATH assessment. In so far as performance on the ELA and MATH assessments do not vary much at low levels of proficiency on the ELP assessment, one might argue that the former assessments are not providing useful information until students reach a point on the ELP assessment where subsequent changes in ELP performance will be associated with changes in ELA or MATH performance. However, this interpretation is strictly dependent on current instruction. The observed relationships among these tests reflect the specific construction of these particular assessments and the way that instruction is approached currently for ELLs. If the language standards of the ELP and content assessments were closely aligned, if the ELP assessment emphasized academic language development at the low end of the scale, or teachers were able to increase their development of students' content-area knowledge regardless of their level of language proficiency, these relationships might reasonably be expected to change. One thing is fairly clear from Figures 1 and 2; namely, regardless of the reason or reasons for the observed increase in variability in ELA and MATH performance at the upper end of the ELP scale, this phenomenon and the general relationship of the ELP and content assessments are both consistent across grades and content areas, and neither appears to vary considerably as a function of Years in U.S.

Not so obvious, due to the arrangement of Figures 1 and 2, is the fact that ELP_SS, ELA, and MATH performance increased fairly steadily as Years in U.S. increased. This relationship holds in grades 4 through 8, although the relationship is somewhat stronger in the earlier grades, and is considerably stronger for ELP_SS than for ELA and MATH. For example, in grade 4, ELP_SS performance increased from 347 to 382 from one to five Years in U.S., whereas ELA increased from 223 to 226 and MATH increased from 220 to 226. While the greater increase for ELP_SS is due in part to the greater variance in ELP_SS (*SD* = 26.71) than ELA (*SD* = 12.19) and MATH (*SD*=14.31), this difference does not fully account for the differential effect of Years in U.S. across the three domains. Measured in standard deviation units, the difference in ELP_SS for students with 1 and 5 years experience in U.S. schools was 35/26.71 = 1.29 as compared to .25 for ELA and .42 for MATH. This same general pattern applies in all grades, although the differences due to Years in U.S. for ELA and MATH were less in grades 7 and 8 and were negligible for MATH in grade 8.

Random Effects Regression Models

To examine the relationships among Years in U.S., ELP performance, and content-area performance, we ran separate random effects regression models for ELA and MATH as

RETURN to TABLE of CONTENTS

Figure 1. Relation between ELA and ELP Assessments by Grade and Years in U.S.

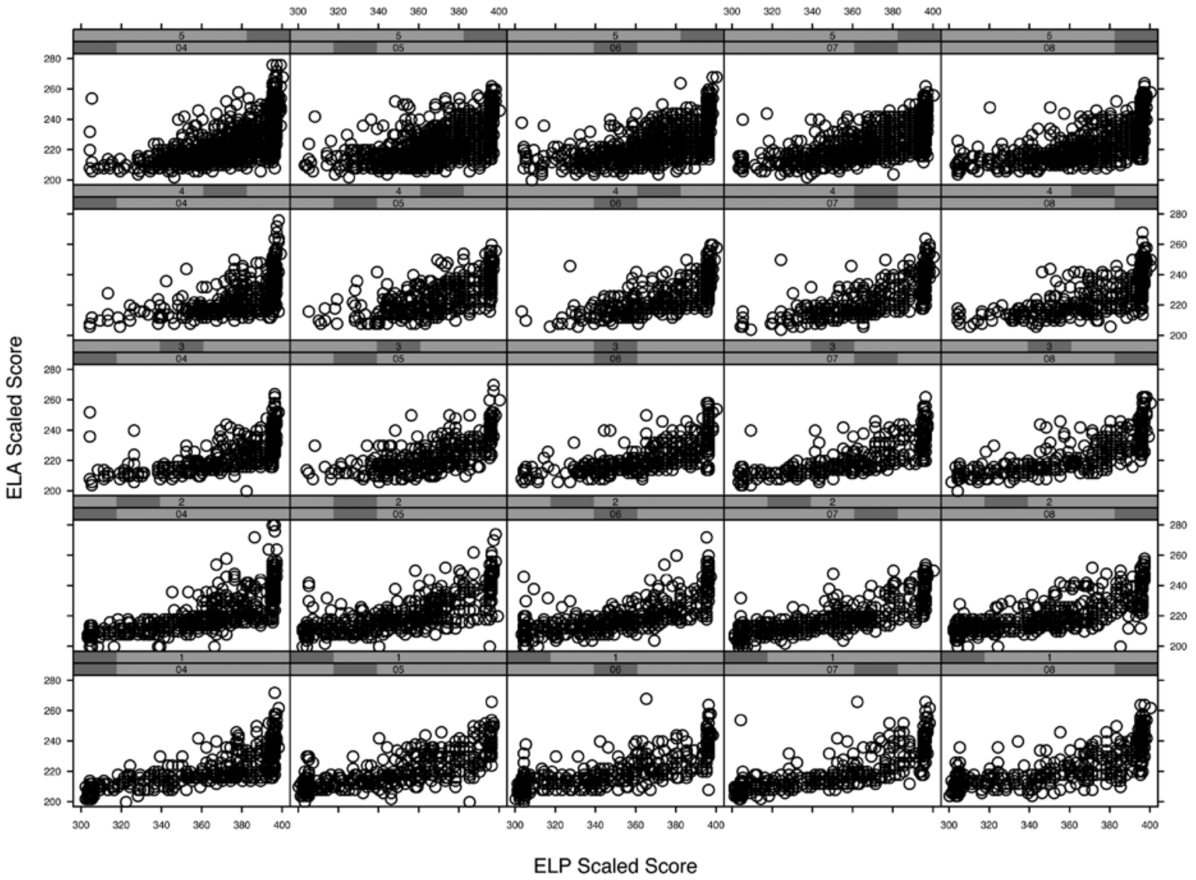
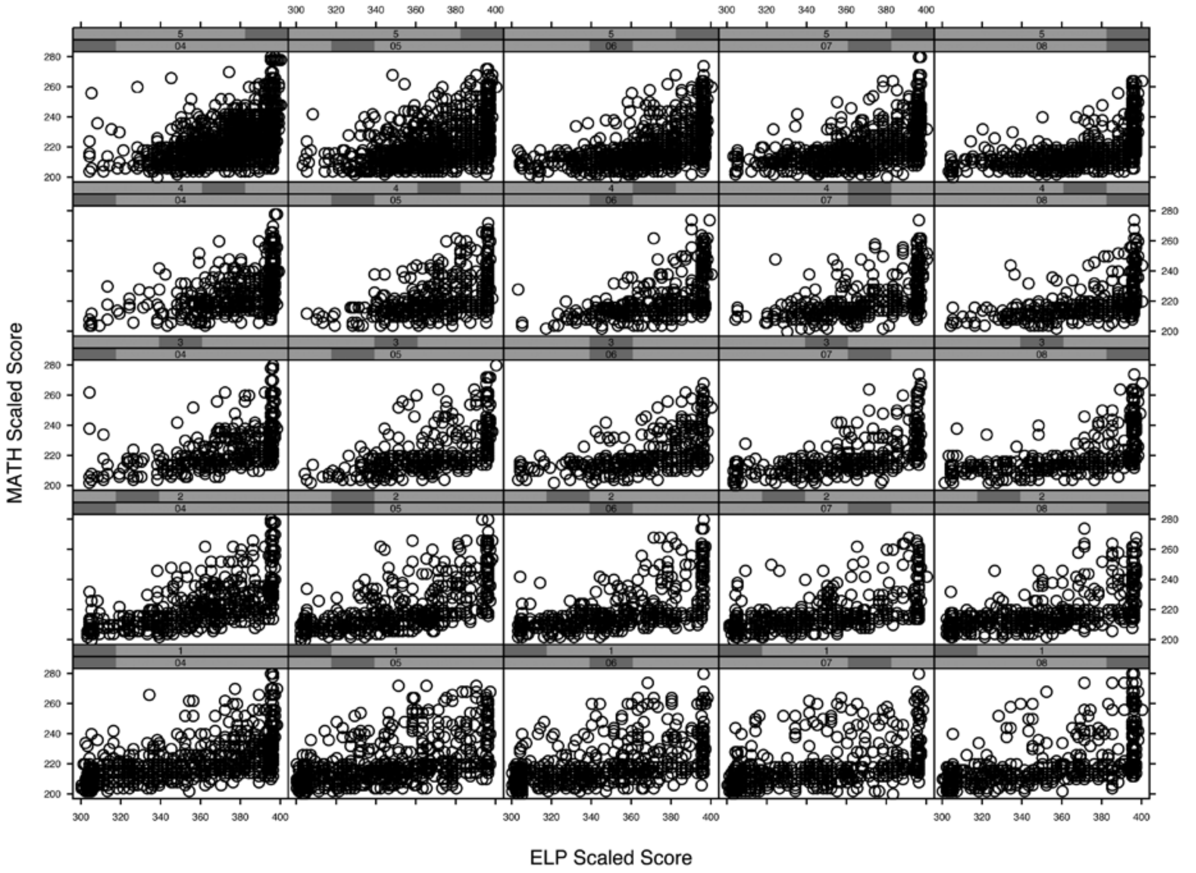


Figure 2. Relation between Math and ELP Assessments by Grade and Years in U.S.



RETURN to TABLE of CONTENTS

described above. Results for these models are presented in Tables 1 and 2. Results for the unconditional models (i.e., models with no predictors) show that significant variability in ELA and MATH outcomes is present at the student, school, and district levels. In general, schools and districts each accounted for 10–20% of the total variability in scores, with somewhat higher values for districts for MATH performance in grades 6–8. Variability due to differences between students within schools ranged from 57% to 72%, indicating that for all grades and content areas, the majority of the variability in scores is between students within schools.

Tables 1 and 2 also present results for models that used Years in U.S. alone and in conjunction with ELP_SS as predictors of ELA and MATH. To simplify presentation, Tables 1 and 2 are organized to show the reductions in variance at the district, school, and student levels that result from adding in the student-level predictors. In multilevel models of this type, there is no guarantee that the variance accounted for will be positive at all levels. Briefly, variance estimates can increase at the school and district level when predictors are entered at the student level, indicating that differences between schools and/or districts are greater when the student-level predictor is controlled.

Examining the results for the conditional models in Tables 1 and 2, it is clear that ELP_SS is the more important of the two predictors. When both ELP performance and Years in U.S. were entered together in the model, the variability dropped substantially at all levels of the models. Anywhere from 40% to 70% of the variance at the school and district levels are accounted for by ELP performance in combination with Years in U.S., most of which is associated with ELP performance. Although the tables show the effect of having both ELP and Years in U.S. in the model, the increment can be judged by comparing the variance estimates for the two different models. In all models, both Years in U.S. and ELP are statistically significant, as are the variance components at all three levels. In the models for MATH, the inclusion of both Years in U.S. along with ELP performance levels makes a more substantial impact in explaining variability at the district level as compared to the models for ELA. For MATH, inclusion of Years in U.S. and ELP increases the variance accounted for by more than 50% in grade 8, and by one third or more in grades 4 and 6, and just under one third in grade 7.

It is interesting to note that only 30% to 40% of the variability in ELA and MATH outcomes at the student level is accounted for by inclusion of the two student-level predictors of Years in U.S. and ELP performance. Even when using both predictors, both of which are statistically significant in the models, these estimates of variance accounted for at the student level are about half the magnitude of those observed in the analysis of the Minnesota data. Recall, however, that the Minnesota analysis did not account for variability at the district and school level. Because districts and schools differ in the language proficiency and achievement levels of their students, these relationships should be examined in a multi-level framework when attempting to estimate the relationship at the student level.

Summary

Taken together, the results for the unconditional and conditional models, as well as the descriptive and exploratory analyses, suggest that it is possible to use the ELP assessments in a meaningful way to index ELLs' progress towards proficiency on the ELA and MATH assessments. Moreover, the results of the conditional models indicate that the key determinant of student performance on the ELA and MATH assessments is performance on the ELP assessment, not Years in U.S. This conclusion is based on the observations that (1) Years in U.S. was less strongly related to performance on ELA and MATH than ELP performance was, (2) that Years in U.S. was more strongly related to ELP development than to ELA and MATH performance, and (3) that models predicting from Years in U.S. and ELP were for the most part comparable to models that used ELP assessments only. These latter models were not shown here in the interest of space, but are available from the first author upon request. Although Years in U.S. remained a statistically significant predictor in the models that included the ELP composite scaled score, the effect of Years in U.S. was not found to be systematic (i.e., means did not increase systematically with Years in U.S. as they did with performance on the ELP assessment), and differences in outcomes for different values of Years in U.S. tended to be on the order of 1 to 3 points when ELP performance was controlled.

ELL Assessment and Accountability

There is widespread concern that the accountability framework of NCLB is not well suited to the needs of ELLs (Abedi, 2004). Unlike other student groups targeted under NCLB, the characteristic of limited English proficiency is considered temporary because the student's membership in

RETURN to TABLE of CONTENTS

Table 1. Estimates of variance components and variance accounted for based on unconditional and conditional models for ELA.

Grade	Source	ELA					
		Unconditional Model	% Var.	Years in U.S.	ΔR^2	Years + ELP_SS	ΔR^2
4	District	23.99	0.15	27.21	-0.13	14.00	0.42
	Schools	25.61	0.16	25.04	0.02	14.90	0.42
	Students	111.69	0.69	108.37	0.03	74.28	0.33
5	District	25.20	0.17	25.73	-0.02	10.72	0.57
	Schools	15.58	0.11	14.83	0.05	8.27	0.47
	Students	107.48	0.72	104.37	0.03	65.37	0.39
6	District	21.05	0.15	22.15	-0.05	7.02	0.67
	Schools	20.47	0.14	18.24	0.11	10.78	0.47
	Students	100.77	0.71	97.03	0.04	61.97	0.39
7	District	25.79	0.17	27.88	-0.08	11.09	0.57
	Schools	17.57	0.12	13.08	0.26	4.05	0.77
	Students	108.15	0.71	104.51	0.03	57.85	0.47
8	District	26.05	0.16	26.70	-0.02	8.36	0.68
	Schools	24.18	0.15	22.99	0.05	7.14	0.70
	Students	115.44	0.70	113.83	0.01	69.24	0.40

Note: ΔR^2 *computed as change in variance component from unconditional model relative to magnitude of variance component in unconditional model. % Var. expresses the variance at each level as a percentage of the total variance.*

Table 2. Estimates of variance components and variance accounted for based on unconditional and conditional models for MATH.

Grade	Source	MATH					
		Unconditional Model	% Var.	Years in U.S.	ΔR^2	Years + ELP_SS	ΔR^2
4	District	36.70	0.17	41.11	-0.12	25.27	0.31
	Schools	34.10	0.15	32.74	0.04	22.08	0.35
	Students	34.10	0.15	32.74	0.04	22.08	0.35
5	District	34.10	0.15	32.74	0.04	22.08	0.35
	Schools	34.94	0.15	33.28	0.05	22.44	0.36
	Students	151.28	0.66	149.57	0.01	112.85	0.25
6	District	48.79	0.23	49.56	-0.02	28.26	0.42
	Schools	23.93	0.11	23.81	0.01	18.99	0.21
	Students	135.55	0.65	133.72	0.01	104.52	0.23
7	District	58.80	0.29	61.72	-0.05	42.19	0.28
	Schools	20.00	0.10	19.42	0.03	14.21	0.29
	Students	120.66	0.60	119.63	0.01	93.23	0.23
8	District	52.35	0.27	51.31	0.02	35.03	0.33
	Schools	29.67	0.15	30.17	-0.02	20.94	0.29
	Students	110.01	0.57	109.00	0.01	85.41	0.22

Note: ΔR^2 *computed as change in variance component from unconditional model relative to magnitude of variance component in unconditional model. % Var. expresses the variance at each level as a percentage of the total variance.*

RETURN to TABLE of CONTENTS

the ELL category changes as the student masters English. In the past, once a state determined that a student had attained English proficiency, they were no longer included in the ELL category in reports on adequate yearly progress, but were moved into the general student category, as well as relevant categories for gender and ethnicity. Because mastery of English is a direct determinant of their mastery of content-area knowledge, the practice of removing students from the subgroup amounts to creaming from the top of the achievement distribution within the subgroup of ELLs, thereby giving a distorted view of how ELLs fare in our educational system in the long run.

This year, the U.S. Department of Education released the new Title I regulations, which address the concern that states, districts, and schools get credit for the progress of ELLs in adequate yearly progress (AYP) determinations. In response, the new regulations permit a state to make AYP determinations by including former ELLs in the ELL category for up to two years after they no longer meet the state's definition of ELL. In addition, the new regulations permit a state to exempt recently arrived ELLs from one administration of the state's reading/language arts assessment (U.S. Department of Education, 2006). However, these provisions really do not go far enough in ensuring an effective accountability system for ELLs. For one, the long-term outcomes of ELLs in the educational system is unavailable because of the temporary nature of their status. The reason for changing the status of students who have gained proficiency is to prevent states from receiving funds for language support services for students who no longer need those services. However, this problem could certainly be solved in a way that allows for more accurate accounting of how ELLs fare in U.S. public schools. One possible solution is to create a category of *Former ELL*, to keep students in this category for reporting purposes throughout their schooling. Whether students should remain in this category as students transfer from one school to another, one district to another, or one state to another could be debated, but the creation of such a permanent category for reporting purposes would be a step in the right direction.

An alternative solution, and one that we think is preferable, would be to designate ELLs for reporting purposes on the basis of their levels of language proficiency. In this model, the achievement results of ELLs should be reported by language proficiency bands. The number of proficiency bands could easily be restricted to four or five, with the top band being fluent English proficient, the category discussed above for students who are no longer ELL. One would expect that content-area achievement results would decline from the highest to the lowest levels of English language proficiency. However, one would also expect that those students in the fluent English proficient band should perform comparably to monolingual English speakers in terms of pass rates on state content assessments. More importantly, it would not be unreasonable to expect that schools could consistently work to improve the content-area achievement results for students within each language proficiency band on a year to year basis as schools become more effective at delivering instruction in linguistically sensitive ways to students at each language proficiency band. Reporting results on achievement tests for ELLs—conditional on levels of language proficiency—would be an improvement over current reporting systems because year to year results would be more sensitive to the effects of instruction when a major determinant of academic performance for ELLs is being held constant in each reporting category. To be complete, the system would also track progress through the language proficiency bands as a function of years in the state, with the expectation that students are individually progressing from less proficient to more proficient categories and that schools continually improve on the rate with which students progress through the language proficiency bands as schools become more effective in developing students' language proficiency per year of instruction. Such progress would be apparent in the percentage of students in each language proficiency band with a given number of years in the state's schools. Such improvements in the distribution of language proficiency conditioned on years in the state would be measurable and actionable by schools, districts, and states.

While these improvements to the accountability system can be made without significant changes to current assessment and accountability systems, they would still lack the ability to motivate teachers and students because it is unclear what the target objective is for a student at lower levels of the language proficiency continuum when they begin school in any given school year. One way to address this issue is through the concept of a developmental index that takes into account the development of language and content-area knowledge and the fundamental relationship between them. There are numerous ways that such an index could be constructed.

One such alternative would be to develop a weighted composite of the language proficiency assessment and the

RETURN to TABLE of CONTENTS

content-area assessment such that the weights vary as a function of the number of years that the student has been in the state. For example, for ELLs with only one year in the state, almost all of the weight would be placed on the outcome of the language assessment, say 90%. Similarly, for ELLs with two years in the state, the weights would shift to place more weight on the content assessment, say 40%, up from 10% in the first year. Over some specified period of time (e.g., four years) the weights would have shifted to the point where 100% of the weight is on the content assessment. Such an index has the advantage of emphasizing to teachers and students that attention needs to be placed both on language development and on the development of content-area knowledge. Such an index further emphasizes the fundamental role that language plays in the development of content-area knowledge. Optimally, the weights should be chosen through statistical analysis of language and content-area assessment data for each state, and could be developed through a series of two-wave longitudinal datasets. The advantage of such an index system is that it is developmentally sensitive, it reflects the confounding role that language development plays in the development of content-area knowledge, and it holds all children to the same long term educational outcome standards. Most importantly, such an index model allows all children to contribute as a success in their school's AYP determination each and every year that they attend that school by meeting a goal that is developmentally appropriate and challenging for them. Moreover, if they make their goal every year, they will be achieving at the same level of academic proficiency as their monolingual English language peers within a specified number of years, but they do not have to wait until then to contribute to the measured success of their school.

Conclusion

Through this chapter we have reviewed some of the history around the role of language testing in U.S. public schools along with the current status of ELLs under NCLB. We have considered some of the academic challenges faced by English language learners. An important element in our success in this endeavor is the continued development and refinement of ELP tests and their increased integration with measures of content-area achievement. The ultimate goal of this increased integration and alignment is improved instruction and ultimately higher levels of achievement for ELLs. To realize these goals, we must continue to press for the development of better tests, improved instruction, and stronger links between the two. At the same time, there are immediate steps that we can take to improve the assessment and accountability systems that we use for monitoring the academic achievement and language proficiency of ELLs and for holding schools, districts, and states accountable for this important and unique subgroup of the student population.

References

Abedi, J. (2004). The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues. *Educational Researcher*, *33*, 4–14.

Abedi, J. (2007). English Language Proficiency Assessment & Accountability under NCLB Title III: An Overview. In J. Abedi (Ed.), *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 3–10). Davis: University of California.

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, *17(4)*, pp.371–392.

August.S.t, D. & Hakuta, K., eds. (1997). I*mproving Schooling for Language-Minority Children: A Research Agenda*. Committee on Developing a Research Agenda on the Education of Limited-English-Proficient and Bilingual Students; Board on Children, Youth, and Families; National Research Council. Washington, D.C.: National Academy Press.

August.S.t, D. L., & Shanahan, T. (2006). *Developing literacy in a second language: Report of the National Literacy Panel*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bailey, A. L., & Butler, F. A. (2007). A conceptual framework of academic English language for broad application to education. In A. Bailey (Ed.), *The Language Demands of School: Putting Academic English to the Test*. New Haven, CT: Yale University Press.

RETURN to TABLE of CONTENTS

Bates, E., Dale, P.S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher and B. McWhinney (Eds.) *Handbook of Child Language* (p. 96–151) . Oxford: Basil Blackwell.

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.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California Department of Education, *Schooling and language minority students: A theoretical framework* (pp. 3–50). Los Angeles: California State University.

Cummings, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera, *Language Proficiency and Academic Achievement*. Avon, England: Multilingual Matters Ltd.

Ellis, R. (1994). *The study of second language acquisition*. New York: Oxford University Press.

Francis, D. J., Rivera, M. O., Lesaux, N., Kieffer, M., & Rivera, H. (2006)(a). *Practical Guidelines for the Education of English Language Learners: Research-based Recommendations for Instruction and Academic Interventions*. Portsmouth: RMC Corporation.

Francis, D. J., Rivera, M. O., Lesaux, N., Kieffer, M., & Rivera, H. (2006)(b). *Practical Guidelines for the Education of English Language Learners: Research-Based Recommendations for the use of Accommodations in Large-Scale Assessments*. Portsmouth: RMC Corporation.

Garcia, P. (2003). The use of high school exit examinations in four southwestern states. *Bilingual Research Journal*, *27(3)*, pp. 431–450.

Grigorenko, E.L., Sternberg, R.J., & Ehrman, M.E. (2000). A theory-based approach to the measurement of foreign language learning ability: The CANAL-F theory and test. *The Modern Language Journal*, *84*, 390–405.

Hakuta, K., Butler, Y. G., and Witt, D. (2000). *How long does it take English learners to attain proficiency?* University of California Linguistic Minority Research Institute, Policy Report 2000-1. http://www.stanford.edu/~hakuta/Docs/HowLong.pdf

Halliday, M.A.K.(1994). *An introduction to functional grammar (2nd ed.)*. London: Edward Arnold.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes' *Sociolinguistics* (Eds.), Harmondsworth, England: Penguin. pp. 269–293.

Kato, K. Albus, D. Liu, K. Guven, K. & Thurlow, M. (2004). *Relationships between a statewide language proficiency test and academic achievement assessments*. (LEP Projects Report 4). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.

Krashen, Stephen D. (1988). *Second Language Acquisition and Second Language Learning*. Prentice-Hall International.

Lau v. Nichols, 414 U.S. 563 (1974).

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

McNamara, T. (2006). Assessment of second language proficiency. In K. Brown (Ed.), *Encyclopedia of language and linguistics 2nd edition* (p. 546–553). Amsterdam: Elsevier.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement*, 3rd Edition. New York: American Council on Education and MacMillan. p. 13–103.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice*, *21(1)*, 30–43.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*, 477–496.

Muñoz-Sandoval, A., Cummins, J., Alvarado, C. G., & Ruef, M. L. (1998). *Bilingual Verbal Ability Test Comprehensive Manual*. Itasca, IL: Riverside Publishing.

National Center for Education Statistics. (2005) *National Assessment of Educational Progress, 2005, reading assessments*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Research Council. (2004). Keeping score for all. Washington, DC: National Academies Press.

Raymond, E. B. (2004). *Learners with Mild Disabilities: A Characteristics Approach (2nd ed.)*. Boston, MA: Allyn and Bacon.

RETURN to TABLE of CONTENTS

Scarcella, R. (2003) Academic English: A Conceptual Framework. UC Linguistic Minority Research Institute Technical Report 2003-1.

Solomon, J., & Rhodes, N. C. (1995). *Conceptualizing Academic Language*. National Center for Research on Cultural Diversity and Second Language Learning. Washington, DC.: Center for Applied Linguistics.

Snow, C., Burns, M., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academy Press.

Snow, C., & Kim, Y. S. (2007). Large problem spaces: The challenge of vocabulary for English Language Learners, in R. K. Wagner, A. E. MU.S.e, & K. R. Tannenbaum (Eds.), *Vocabulary Acquisition: Implications for Reading Comprehension*. New York: The Guilford Press. p. 124.

Stahl, S., & Fairbanks, M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56(1)*, 72 –110.

Stoynoff, S. & Chapelle, C.A. (2005). *ESOL Tests and Testing*. Alexandria, VA: TESOL.

Thomas, W.P. & Collier, V. P. (2001). *A National Study of School Effectiveness for Language Minority Students. Long-Term Academic Achievement*, Center for Research on Excellence and Diversity in Education (CREDE), available online at http://www.crede.ucsc.edu/research/llaa/1.1_final.html. (Retrieved in PDF format on 8/28/2002).

U.S. Department of Education (2006). *New* No Child Left Behind *Regulations: Flexibility And Accountability For Limited English Proficient Students*. Press release: September 11, 2006.

U. S. Government Accountability Office (2006). *No Child Left Behind Act: Assistance from Education Could Help States Better Measure Progress of Students with Limited English Proficiency*. Report to Congressional Requesters.

Valdés, G., & Figueroa, (1994). *Bilinguism and Testing: A Special Case of Bias*. Norwood, NJ: Ablex Publishing Corporation.

Ziegler, J. & Goswami, U. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131(1*), 3 –29.

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

Chapter 3

Developing the Mountain West Assessment

Ginger Mathews

The Mountain West Assessment Consortium (MWAC) received a two-year Enhanced Assessment Grant awarded by the U.S. Department of Education under Title VI, Subpart I, Section 6112 in 2003 for which the Utah State Office of Education (USOE) served as the fiscal agent. The consortium partners were a group of states located primarily in the mountain west and northern plains regions and Measured Progress, a not-for-profit educational assessment organization. Specifically, the consortium partners included state departments of education in Alaska, Colorado, Idaho, Michigan[1], Montana, Nevada, New Mexico, North Dakota, Oregon, Utah, and Wyoming[2].

Introduction

The No Child Left Behind Act of 2001 (NCLB) was designed to improve performance of the country's elementary and secondary schools. NCLB includes increased accountability for states, districts, and schools through challenging state standards in reading and mathematics, testing at both the elementary and secondary school levels, and creating annual statewide progress targets aimed at the goal of all students reaching proficient status by 2014. As part of reaching the goals of student proficiency and improved school performance, under NCLB, states are required to show that English language learners (ELLs) are demonstrating improvements in both English proficiency and academic content. Under NCLB, Title III: *Language Instruction for Limited English Proficient and Immigrant Students* focuses on providing the services and tools for English learners to meet the same challenging state standards required for all students. As part of Title III, annual assessment of students' language proficiency is required. Historically, these types of assessments have focused on basic language skills, rather than academic language (the language of the classroom and content areas). The Mountain West Assessment Consor-

[1] Although not a mountain west or northern plains state, Michigan requested to join the project in the second year of the grant and became a full member along with the other states listed above.

[2] Wyoming participated in the grant for the vast majority of this project, but withdrew in the final few months. We list Wyoming as a participant to recognize the contributions of Wyoming staff to the final products.

RETURN to TABLE of CONTENTS

tium was formed with the goal of developing assessment tools that would measure academic language proficiency.

The focus of the consortium was to improve the assessment and instruction of ELLs given the inability of traditional language proficiency tests to predict the readiness of ELL students to function independently in mainstream English language classes. For this group of states, most ELLs are of Hispanic or Native American cultural/language heritage, although there are many other language groups represented. The wish of the states was to provide enhanced assistance in acquiring proficiency in English, the language of mainstream classrooms in the United States.

In the consortium states, when academic subject test results and other accountability data are disaggregated by ethnicity, larger numbers of ELLs are found to regularly perform below native English language speakers, despite states' considerable efforts to provide appropriate instructional programs. Reasons for this consistent difference in student academic performance can be related to students' lack of familiarity with academic English (English language skills required for the academic context of the mainstream classroom) and lack of available language proficiency assessments that adequately measure academic English, the language of the classroom. In the education of ELLs, it is becoming more common to find two teachers in a classroom, one primarily for English instruction and another who can help English language learners. This practice is being employed to give ELL students a better opportunity of learning and understanding academic subjects and the type of English used to teach them (Zehr, 2006). The consortium agreed that the development of effective classroom-based English language proficiency assessments for ELLs in this region would lead to more appropriate English language and academic instruction and, in turn, to better education and enhanced life opportunities.

In some efforts to identify the most effective ways to teach and test ELL students, researchers have compared students who take an English version of a test versus a Spanish version (Abelle, Urrutia, Shneyderman, 2005). Some findings show that students perform better taking a test in their home language. More importantly, comparisons such as these also show that giving ELL students an English version of a test does not accurately assess their knowledge of the subject matter.

For many ELLs, the lack of language proficiency (in some cases, in both their home language and in English) has limited the extent to which they have experienced success in school. The result is lack of progress in acquiring academic English—as documented on assessments such as the National Assessment of Educational Progress—and a disproportionately high dropout rate. This lack of progress record translates into diminished life opportunities.

The goal of the Mountain West Assessment Consortium was to begin to break this cycle by better identifying the language proficiency of students upon entry into school and better assessing the progress that students were making in learning English in realistic academic contexts. English as a Second Language (ESL) testing has generally been used to rank students according to language abilities. The proposed assessment system would focus on language proficiency growth in order to be a tool to improve learning, rather than a tool for selection, identification, tracking, and sorting. This instrument would set new precedents in the manner in which ELLs are assessed since most existing instruments for the assessment of English language proficiency focus on social English language skills rather than the academic English language skills required for performing well in the mainstream classroom in the United States.

An often-cited problem with current ESL tests is their failure to recognize and assess academic language. As pointed out by Cummins (1984), most ESL tests focus on basic interpersonal communicative skills (BICS) and ignore cognitive academic language proficiency (CALP). CALP refers to the type of language that is learned in the classroom, and is used to deal with the various disciplines. The failure of ESL tests to address CALP is a major limitation when they are used in an assessment system that focuses on the mastery of discipline-based content standards. It is also limits the validity of these tests when they are used for reclassification purposes (to determine student readiness for mainstream classrooms). The goal of MWAC was to address the need for improved academic language instruction and assessment by advancing the field through the creation of a new type of assessment for ELLs.

English Language Development Standards

Upon the award of the grant, the states in the consortium worked with Measured Progress and other consultants to develop a set of common English language development (ELD) standards, later referred to as the Fountain Document. MWAC members and staff from Measured Progress collected the ELD standards from each state. Since many

RETURN to TABLE of CONTENTS

of the states' ELD standards were under development, the consortium agreed to use the Colorado standards as a starting reference point for the development of MWAC ELD standards. The ELD standards were reviewed by external consultants, who analyzed the state ELD frameworks to determine areas of consistency and/or inconsistency with the goal of creating a consensus framework. Of course, the MWAC ELD standards were organized by the four language skill domains (listening, reading, speaking, writing), and also, as is typical of ELD standards, by proficiency levels. (That is to say, language skills are commonly sorted by levels of sophistication. MWAC used three levels initially: *beginning*, *intermediate*, and *advanced*). The Foundation Document (MWAC ELD standards) were used to guide item development.

Title III guidelines require that comprehension be assessed along with the four language skill domains. The MWAC participants chose not to include this category in its blueprint, reasoning that comprehension is clearly something that would be measured in conjunction with reading and listening. Furthermore, it was acknowledged that within each language mode and proficiency level, a student's critical thinking skills; literary response and expression; and linguistic, sociolinguistic, and sociocultural competence would be assessed.

Because of the common criticism that existing language proficiency tests addressed social communications and were not predictive of students' ability to function independently in English-speaking classrooms, MWAC participants decided that the language skills identified in the ELD standards would be assessed in the context of basic English language arts, mathematics, and science academic content standards that were common to the MWAC states (the "content commonalities"). For this reason, the consortium members and contractor reviewed the member states' content standards in the different subject areas. The purpose of this was to identify topics for test material (e.g., reading passages, scripts for listening activities, etc.) that were "fair game" for use as contexts. So that content knowledge would be less of a confounding issue in the language proficiency test, topics for contexts were chosen from content standards for grades below the target grades for the language proficiency tests that were to be developed.

Test Blueprint & Design

Test Blueprint

The consortium created a Test Design Subcommittee to define the test framework. This blueprint was then presented to the full consortium to be approved. The proposed general test blueprint for the completed assessment, drafted by the consortium early in the project, is provided in Table 1.

Table 1 shows the importance that the consortium states attached to the use of content area standards for the MWAC English language proficiency tests. Even though federal reporting requires three levels of English proficiency (*entry-level, intermediate, advanced*), the consortium decided to report at five levels (*pre-emergent, emergent, intermediate, fluent,* and *advanced*). The decision was based on mirroring the levels of language learners (*pre-emergent, emergent, intermediate, fluent*), with the *fluent* level needed for Title III Adequate Yearly Progress (AYP) reporting, and for evaluating program effectiveness.

Following the development of the general test blueprint/framework, the consortium and Measured Progress specified the number and types of items, the specific

Table 1. MWAC English Language Proficiency Test Blueprint

English Language Proficiency Level	English Language Arts	Mathematics	Science	Social Studies
	*L, S, R, W, C	L, S, R, W, C	L, S, R, W, C	L, S, R, W, C
Pre-Emergent				
Emergent				
Intermediate				
Fluent				
Advanced				

Note: L, S, R, W, C stand for listening, speaking, reading, writing, and comprehension, the required language skills to be assessed under Title I and Title III. However, comprehension would be addressed as part of listening and reading.

RETURN to TABLE of CONTENTS

language skills to be assessed, and the ways to incorporate academic content into the assessment. The goal was to develop a fairly detailed description of the work to be carried out so that states could begin to plan the scope of and identify participants for the item development workshops.

Test Design

The test design was developed based on information gathered from states' experiences with their current ELL assessments. Under consideration were test administration, available data, test indication of language proficiency, research, and the opinions of English language acquisition experts. The test design was later revised based on usability information gathered during the pilot test. Information gathered from teacher surveys on test materials, administration time, student experience, and overall ease of administration guided the revision of the test design.

The original test design was organized so that at each grade level, each language skill domain could be administered to students at either Level A or Level B. The levels were based on the continuum of language acquisition. Level A of the assessments included items from the early acquisition and intermediate levels of proficiency. Level B included items from intermediate and transitional levels. Using a *locator tool* for a specific grade span, administrators could determine which level of the test should be administered to a particular student. The locator tool was designed as a Likert scale questionnaire.

The test design used for the small-scale spring 2004 pilot is shown in Table 2. For all grade spans, all domains were individually administered, except for reading, writing, and listening tests for Level B in grade spans 3–5, 6–8 and 9–12, where the assessment was administered to groups.

For the fall 2004 field test, a new test design was developed based on feedback from the pilot administration. The revised test design—also used for the operational assessment—allowed for more group administration, which lessened the total testing time for each student and limited the time required for individual administration. In this design, the only individually administered parts of the assessment were the speaking test for all grade spans and all domains in grade span K–1. The revised test design used for the field test and operational assessment is shown in Table 3. (Specific information on the field test is provided in the section titled *Pilot and Field Tests*). Feedback received from the field test confirmed that the revised test design lessened the testing time required in administering the assessment.

Table 2. Spring 2004 Pilot Test Design

	Reading	Writing	Listening	Speaking
K–1		Checklist	A	A
			B	B
1–2	A	A	A	A
	B	B	B	B
3–5	A	A	A	A
	B	B	B	B
6–8	A	A	A	A
	B	B	B	B
9–12	A	A	A	A
	B	B	B	B

Note: Group administered assessments are shaded.

Item Development

Once the plans for the assessments (blueprints, item allocations, test design) had been created, the consortium developed the passages, stimuli, items, and graphics that would be used in the assessment. This section describes the item development process used for the grant. During the test blueprint and design development process, it was decided that multiple item types would be used to assess students. Multiple-choice, short-answer, and constructed-response items were developed, as appropriate for the language skill domain and grade span.

Item development involved local specialists and educators from each member state as item writers. The consortium hosted centrally located item writing workshops, each focused on a specific language skill domain. This allowed for more participation from the states in developing items

Table 3. Fall 2004 Field Test and Operational Test Design

	Reading	Writing	Listening	Speaking
K-1		Checklist		
1-2	A	A		
	B	B		
3–5	A	A		
	B	B		
6–8	A	A		
	B	B		
9–12	B	B		
	A	A		

Note: Group administered assessments are shaded.

RETURN to TABLE of CONTENTS

for all language skill domains and grade spans, as well as offering more opportunities for professional development and access to assessment and bilingual specialists. Educators from all the member states participated in workshops that were held in three of the states.

Item writers for the workshops were selected based on the following criteria:

- Current or former K-12 educators who were:
  - o Bilingual endorsed, or
  - o TESOL endorsed, or
  - o Experienced in instructing Native American/ Alaska Native students, or
  - o Experienced in instructing immigrant/migrant students, or
  - o Bilingual/Special Education endorsed, and/or
  - o TESOL endorsed and reading, writing, math, science content educators

The item-writing workshops were designed so that participants received a half day of item development training followed by two full days of intense item development. The training portion of the workshops focused on how to develop items and how to incorporate academic language skills into the assessment. Item writers became familiar with the consensus lists of learning activities and technical vocabulary—defined by the consortium—and the test blueprints.

After the item-writing workshops, the items were reviewed and refined by Measured Progress and other ELL consultants. Each reviewer was asked to review the items based on alignment to the ELD standards, appropriateness of the context, style consistency, language clarity, grammatical issues, and general readability, as well as ensuring that every item met psychometric conventions (i.e., clear and parallel options). Scoring guides were also reviewed for open-response items. The items were then reviewed by consortium members and state educators before the pilot and field tests. These reviewers were also asked to review the items for alignment and grade-level appropriateness. Based on comments by all reviewers, the items were revised by Measured Progress staff before being used in the pilot and/or field tests.

Bias and Sensitivity Review

As part of the item development process, all stimuli, graphics, and items were provided to state-selected participants for bias and sensitivity review. Reviewers were asked to read/review graphics, passages, and items and identify any potential for bias or sensitivity issues. For this assessment, bias was defined as the presence of some characteristic of a passage that results in differential performance for two individuals of the same ability but from different ethnic, gender, cultural, or religious groups. Sensitivity issues, such as offensive language, stereotyping, and disturbing topics (e.g., death, family relationships), could lead to bias. During this review, items and stimuli were flagged by participants, and possible bias and/or sensitivity issues were noted. Much attention was paid to the various ethnic and cultural backgrounds of the students who would be assessed using the Mountain West Assessment.

After the review, any items or stimuli that were flagged as having potential for bias or sensitivity issues were reviewed by the lead consortium members and Measured Progress development staff. When possible, items and/or stimuli were revised, redrawn, or rewritten to eliminate the noted concerns. If "re-working" the item was not possible, items were removed from the item pool and future use on the assessment. All final decisions on which items were revised or removed were made by MWAC and Measured Progress.

Pilot and Field Tests

Pilot Test

Following the item development process and review, items were selected to be used in the pilot test and later in the field test. The purpose of the pilot test was to learn more about the assessment design and administration process. The later field test was used to learn about how the items functioned by collecting and analyzing item-specific data.

The design of the pilot test required the following materials:

- One form of each language skill domain, in Level A and Level B
- Compact disc (CD) for listening domain
- Teacher guides (including scoring parameters for each open-ended item)
- Locator test/protocol
- Answer document
- Feedback questionnaire

RETURN to TABLE of CONTENTS

To meet the pilot sample requirements, each state identified about 50 *pre-emergent/beginning* students and 50 *intermediate/advanced* students in each of the five grade spans. Each student was assessed on all four language skill domains. Thus a total of approximately 500 students were targeted from each state. Consortium members worked within their own states to identify participants for the pilot test, thus the sample was not randomly selected.

Measured Progress prepared all testing materials needed for the pilot test. Test booklets, examiner manuals, tally sheets, and the teacher questionnaire were designed, prepared, printed and shipped by Measured Progress. During the pilot test administration, Measured Progress staff provided support to participants by both phone and email.

At the conclusion of the pilot test, each examiner (test administrator) was asked to complete a questionnaire. The questionnaires were analyzed along with other anecdotal data that was collected during and after the pilot test administration. Table 4 summarizes the data collected, concerns that were raised, and the proposed design revisions that were presented to the consortium. Items were refined based on the examiners' responses to the questionnaire (summarized in Table 4), and item review committee comments (as described in the section Item Development), and prepared for use in the fall field test.

Field Test

Following the March 2004 pilot test, the assessment design and items were refined based on the comments received by administrators and content specialists. Adjustments, revisions, and modifications were made to the items, test forms, and test materials. The resulting assessment materials were used in the fall 2004 field test. The field test of the Mountain West Assessment was used to gather information on the revised test design, the locator tool, and item performance. The field test materials were formatted to match the test design established for the operational (final) test forms, except that Levels A and B were merged into one continuous form. This was done to gather information on the appropriateness of the item order and assumption of difficulty.

The original design of the field test used a stair-step model, where each form contained blocks of different items. However, due to a smaller number of participants than anticipated from across the consortium states, the design of the field test and scaling plans were revised.

In the final field test design, four forms were developed for each language skill domain for each grade span. Within each form, the items were positioned by assumed difficulty, according to the test blueprint; no division of Level A and Level B items was made in the test booklet or other testing materials. The total number of items field tested is shown in Table 5. The field test sampling design included both ELL students and native English speakers (NES). The NES students were included in the field test to assess whether the knowledge and skills that were asked of ELLs were indicative of the knowledge and skills possessed by native English speakers. The information obtained from native speakers on the field test items was used to help in the construction of operational test forms.

Field test materials were designed to be the same as those intended for use in the operational assessment. All test materials (test booklets, examiner manuals, answer

Table 4. Pilot Test Concerns and Proposed Test Design Revisions

Pilot Test Concerns	Suggested Revisions and Rationale
Locator tool ineffective	• Eliminate poorly differentiating items • Cut scores needed that will maximally differentiate between Level A and Level B students
Test administration awkward	• Convert to 100% group administered tests for grade spans 1–2, 3–5, 6–8 and 9–12, except for speaking • Revise the specific sequence of activities in pilot instruments to allow for smooth transitions between items
Appropriateness	• Address concerns about specific items having a "middle-class" feel during a full bias and sensitivity review • Address concerns about non-native content sensitivity
Tests administration time too long	• Reduce the number of items by 20 to 25% • Reduce passage/stimulus length • Convert individual administration to group administration were noted above • Restructure listening and speaking domains for all grade spans to a single level

RETURN to TABLE of CONTENTS

Table 5. Fall 2004 Field Test —Total Number of Items per Form and Domain

	K–1		1–2		3–5		6–8		9–12	
	Form	Total	Form	Total	Form	Total	Form	Total	Form	Total
Writing	Checklist		21	84	19	76	19	76	19	76
Listening	22	88	22	88	22	88	22	88	22	88
Speaking	14	56	14	56	14	56	14	56	14	56
Reading	36	144	30	120	29	116	31	124	31	124
	Total	288	Total	348	Total	336	Total	344	Total	344

Table 6. ELL Population Estimate

	ELL Population Estimate	Field Test Target	Field Test Actual (including ELL & NES)
AK	20,057	1,350	1,295
CO	83,824	5,100	5,237
ID	18,746	1,350	1,594
MI	60,876	4,250	3,194
MT	7,043	550	315
ND	6,205	450	1,815
NM	65,259	4,250	1,999
NV	50,000	3,050	394
UT	38,543	2,450	564
WY	3,378	375	0

documents, locator tools, and listening CDs) were prepared, printed, and shipped directly to the field test participants by Measured Progress.

As noted above, the total participation for the field test was considerably smaller than originally expected. In Tables 6 and 7, the ELL population estimates, original goals for participation, and actual number of students tested are shown. After the field test window closed, field test participants returned all test materials to Measured Progress using the supplied return materials. Once the test materials were returned to Measured Progress, they were logged in, scanned, and scored. After the scanning and scoring were complete, electronic files were prepared for item analysis. During this time, analysis on NES student abilities, the locator tool, and item difficulty were conducted, with results to be used in developing the operational forms.

Table 7. Field Test Target and Actual Participation Numbers

	K–1			1–2			3–5			6–8			9–12		
	Target	Actual		Target	Actual		Target	Actual		Target	Actual		Target	Actual	
	ELL	ELL	NES	ELL	ELL	NES	ELL	ELL	NES	ELL	ELL	NES	ELL	ELL	NES
AK	150	174	41	300	90	41	300	282	67	300	239	49	300	233	79
CO	800	507	155	1,075	525	156	1,075	1,044	315	1,075	1,030	270	1,075	1,039	196
ID	150	297	83	300	79	26	300	375	138	300	314	139	300	97	46
MI	650	391	105	900	321	112	900	590	186	900	573	182	900	568	166
MT	50	2	0	125	26	7	125	106	10	125	37	41	125	74	12
NM	50	233	38	100	161	40	100	396	73	100	462	70	100	293	49
NV	650	318	111	900	194	67	900	459	143	900	308	95	900	242	62
ND	450	89	19	650	30	12	650	103	24	650	37	28	650	38	14
UT	350	31	8	525	56	6	525	103	15	525	129	47	525	150	19
WY	35	0	0	85	0	0	85	0	0	85	0	0	85	0	0
Total	3,335	2,042	560	4,960	1,482	467	4,960	3,458	971	4,960	3,129	921	4,960	2,734	643

RETURN to TABLE of CONTENTS

Table 8. Average Difficulty and Discrimination of Different Item Types – Grade Span K–1

	Statistics	Item Type		
		All	Multiple Choice	Open Response
Listening	Difficulty	0.38 (0.19)	0.27 (0.19)	0.44 (0.17)
	Discrimination	0.47 (0.14)	0.42 (0.09)	0.50 (0.15)
	N	87	28	59
Speaking	Difficulty	0.72 (0.16)	--	0.72 (0.16)
	Discrimination	0.55 (0.13)	--	0.55 (0.13)
	N	56	--	56
Reading	Difficulty	0.55 (0.17)	0.57 (0.16)	0.54 (0.17)
	Discrimination	0.48 (0.14)	0.46 (0.11)	0.50 (0.16)
	N	144	72	72

Note: Numbers in parentheses are standard deviations.

Table 9. Average Difficulty and Discrimination of Different Item Types – Grade Span 1–2

	Statistics	Item Type		
		All	Multiple Choice	Open Response
Listening	Difficulty	0.75 (0.18)	0.75 (0.18)	--
	Discrimination	0.36 (0.16)	0.36 (0.16)	--
	N	88	88	--
Speaking	Difficulty	0.80 (0.14)	--	0.80 (0.14)
	Discrimination	0.53 (0.12)	--	0.53 (0.12)
	N	56	--	56
Reading	Difficulty	0.72 (0.19)	0.72 (0.19)	--
	Discrimination	0.43 (0.16)	0.43 (0.16)	--
	N	118	118	--
Writing	Difficulty	0.73 (0.19)	--	0.73 (0.19)
	Discrimination	0.50 (0.16)	--	0.50 (0.16)
	N	84	--	84

Note: Numbers in parentheses are standard deviations.

Summary statistics of the difficulty and discrimination indices for each item are provided in Tables 8 through 12. In general, the item difficulty and discrimination indices are in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall. There were a small number of items with near-zero discrimination indices, but none was reliably negative. Occasionally, items with less-desirable statistical characteristics need to be included in assessments to ensure that content is appropriately covered, but there were very few such cases on the Mountain West Assessment.

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. However, one can say that difficulty indices were fairly similar across four of the five grade spans. For grade span K–1, the difficulty indices tended to be lower (indicating lower performance) than

RETURN to TABLE of CONTENTS

Table 10. Average Difficulty and Discrimination of Different Item Types – Grade Span 3–5

	Statistics	Item Type		
		All	Multiple Choice	Open Response
Listening	Difficulty	0.64 (0.18)	0.64 (0.18)	--
	Discrimination	0.40 (0.12)	0.40 (0.12)	--
	N	88	88	--
Speaking	Difficulty	0.84 (0.11)	--	0.84 (0.11)
	Discrimination	0.58 (0.15)	--	0.58 (0.15)
	N	56	--	56
Reading	Difficulty	0.62 (0.18)	0.62 (0.18)	0.47 (0.15)
	Discrimination	0.44 (0.11)	0.44 (0.11)	0.58 (0.07)
	N	114	110	4
Writing	Difficulty	0.75 (0.16)	0.74 (0.14)	0.75 (0.18)
	Discrimination	0.49 (0.16)	0.48 (0.12)	0.51 (0.19)
	N	74	34	40

Note: Numbers in parentheses are standard deviations.

Table 11. Average Difficulty and Discrimination of Different Item Types – Grade Span 6–8

	Statistics	Item Type		
		All	Multiple Choice	Open Response
Listening	Difficulty	0.78 (0.14)	0.78 (0.14)	--
	Discrimination	0.45 (0.14)	0.45 (0.14)	--
	N	86	86	--
Speaking	Difficulty	0.83 (0.13)	--	0.83 (0.13)
	Discrimination	0.56 (0.16)	--	0.56 (0.16)
	N	56	--	56
Reading	Difficulty	0.66 (0.18)	0.67 (0.17)	0.46 (0.15)
	Discrimination	0.40 (0.13)	0.38 (0.13)	0.57 (0.08)
	N	122	114	8
Writing	Difficulty	0.70 (0.19)	0.67 (0.20)	0.73 (0.18)
	Discrimination	0.47 (0.18)	0.36 (0.13)	0.59 (0.14)
	N	76	40	36

Note: Numbers in parentheses are standard deviations.

those for the other grade spans; the one exception was for open-response reading items, for which the K–1 difficulty indices were higher than those for the other grade spans.

Comparing the difficulty indices of multiple-choice and open-response items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that, in most cases, the difficulty indices for multiple-choice items tend to be higher than the difficulty indices for open-response items. The one exception was for K–1 listening, for which the difficulty value for the multiple-choice items was substantially lower than that for the open-response items. Similarly, the partial credit allowed for open-response items is advantageous in the computation of item-test correlations, so the discrimination indices for these items tend to be larger than the discrimination indices of multiple-choice items.

Differential item functioning (DIF) analyses provide a statistical index that identifies items that may be biased

RETURN to TABLE of CONTENTS

Table 12. Average Difficulty and Discrimination of Different Item Types – Grade Span 9–12

	Statistics	Item Type		
		All	Multiple Choice	Open Response
Listening	Difficulty	0.75 (0.13)	0.75 (0.13)	--
	Discrimination	0.44 (0.12)	0.44 (0.12)	--
	N	88	88	--
Speaking	Difficulty	0.76 (0.20)	--	0.76 (0.20)
	Discrimination	0.55 (0.17)	--	0.55 (0.17)
	N	56	--	56
Reading	Difficulty	0.64 (0.19)	0.65 (0.19)	0.46 (0.14)
	Discrimination	0.42 (0.13)	0.41 (0.12)	0.60 (0.06)
	N	122	114	8
Writing	Difficulty	0.67 (0.18)	0.69 (0.19)	0.64 (0.15)
	Discrimination	0.45 (0.15)	0.41 (0.14)	0.52 (0.13)
	N	75	47	28

Note: Numbers in parentheses are standard deviations.

against particular subgroups. Because of the relatively small number of students tested for the MWAC field test, however, it was not possible to run DIF analyses. Qualitative checks for bias and sensitivity were completed through committee reviews of all items to be piloted in spring 2004. These checks applied to passages, items, and graphics to remove possible differential performance for individuals with the same ability but from different ethnic, gender, cultural, or religious groups.

Operational Form Development and Production

Following the analysis of the field test data, Measured Progress staff began the task of selecting and constructing the operational test forms. Three final forms were developed. As decided by the member states, two secure operational forms were developed as print-ready forms, with the third form provided in the format of an item bank. It was also decided that all item content and data, graphics, and scoring information be provided electronically to each member state for future use and development.

As a result of the low participation numbers in the field test, it was not possible to employ a design that enabled the pre-equating of operational test forms based on field-test analyses. Instead, overlapping or equating items were included in each of the two secure forms to aid the future analysis of students' abilities and performance.

Each test form was developed based on the test blueprints used for the field test. Based on the item analysis data collected from the field test on both ELL and NES responses, items were eliminated from the operational item-pool if they were deemed too difficult or did not function well on the field test. Statistics were reviewed item by item. Item placement and the test blueprints were also revised slightly based on the actual difficulty of items determined by the field test, so that the item order on the assessments appropriately reflected a definite progression based on student proficiency. The final set of forms included Level A and Level B forms where appropriate. The locator tool cut points were also finalized based on field-test data.

Measured Progress developed the final assessments using its standard production procedures and tested processes and procedures. Each test form passed through several quality control and editorial review steps before being considered final and ready for print.

The operational test materials for each form included test booklets, an examiner manual, a scoring manual, a listening CD, a locator tool, and answer sheets (where applicable). The materials were designed to be easy to use and duplicate, based on each state's needs. Much of the design of the final forms was based on the feedback from the bias and sensitivity reviews and information collected

RETURN to TABLE of CONTENTS

after both the pilot and field tests. At the conclusion of the development and production process, the Mountain West Assessment Consortium became the sole owner of all test materials related to the Mountain West Assessment.

Expert Panel: Standard Setting

Formal proficiency-level cut scores were not determined for the MWAC English proficiency assessments, primarily because the assessments had not been administered operationally in any state, and because adopting proficiency expectations would be a state policy decision. Although it was agreed that the project would not propose actual proficiency cut scores, it was desirable to provide guidance to states that could be useful when it was time for each state to set such standards. In an effort to provide this guidance, an expert panel to provide recommendations was born.

To this end, Measured Progress convened a group of national experts in English language acquisition and formed the MWAC Expert Panel Regarding Proficiency Levels to recommend cut scores for state panels. The Leadership Team and Measured Progress believed it would be useful to states to have a recommendation from a group of experts regarding where states might begin their state discussions about cut scores. Similar to recent initiatives in setting standards for state large-scale general assessment programs in which panelists have been brought together for validation studies, the existence of proposed cut scores on the Mountain West Assessments would allow states to bring together a group of standard setters for a validation study. The Expert Panel activities were not meant to replace a formal, within-state standard setting.

Outcomes of the Expert Panel Meeting

Using a modified-bookmark method for standard setting, the Expert Panel produced two types of outcomes: starting point recommendations for state standard-setting discussions and advice to states about setting standards.

Starting Point Recommendations

The Expert Panel recommended proposed starting points for two cut scores for grade span 3–5 and two cut scores for grade span 9–12. At both grade spans, the cut scores involved were the emergent/intermediate and fluent/advanced cut scores. Using an equipercentile smoothing technique, an average of the percentage of students at or above each cut score was taken and applied to all grade spans. The same equipercentile procedures were repeated for the emergent/intermediate cut score. The resulting proposed cut scores are contained in Table 13. The information provided in the table shows the percent of students who would be at or below each cut point.

Advice to States

In addition to these proposed cut scores, the panel provided advice to states regarding setting proficiency standards. In general, the advice took the form of cautions about over reliance on the proposed starting points and suggestions for next steps. The panelists advised the states as follows:

- Due to the very challenging nature of proposing cut scores on an ordered item booklet containing all four language skill domains,
  - states may want to consider setting proficiency levels separately by language skill domain,
  - however, states then must grapple with how to combine reading, writing, listening, and speaking results for each student into an overall composite score.
- Expand the MWAC draft proficiency-level descriptors; the existing ones are too general to serve multiple grade spans

Table 13. Percent of Students Below Cut Points

	K–1	1–2	3–5	6–8	9–12
Emergent/Intermediate Average Cut	10.4	10.4	10.4	10.4	10.4
Emergent/Intermediate Panel Recommendation			13.9		6.9
Fluent/Advanced Average Cut	96.3	96.3	96.3	96.3	96.3
Fluent/Advanced Panel Recommendation			97.7		95.0

RETURN to TABLE of CONTENTS

- States should collect their own empirical data for standard setting.
- Select standard-setting participants who have both English language acquisition training and grade-level experience.

Dissemination to the Consortium

The MWAC Leadership Team met in December 2005. The primary purpose of this meeting was to debrief from the Expert Panel meeting and provide state members with professional development regarding standard-setting methods.

Measured Progress and the Utah State Office of Education representative, who participated in the Expert Panel, reiterated the intended purpose of the activity and discussed the usefulness of the outcomes. Consortium members were informed that the order of item difficulty was likely to change to some extent, based on live administration within their own states. These limitations were outlined in order to emphasize the importance of using the Expert Panel's recommendations as starting points to begin conversations within each state.

In an effort to familiarize the Leadership Team with standard-setting techniques, part of the December 2005 meeting was spent walking through the Expert Panel's activities.

In recognition that states would later need to set proficiency standards on the Mountain West Assessments, Measured Progress psychometricians delivered an informational presentation to consortium members that outlined the steps that states would typically go through to conduct a successful standard-setting meeting, including the fundamental step of selecting one or more standard-setting methods.

Accommodations

The scope of work for the MWAC grant did not include specific collection of data or research on the use of accommodations, once the Mountain West Assessment was revised to meet the consortium and timeline requirements. For the pilot and field tests, recommendations were made for examiners to use their state's standard accommodations as needed for their students. MWAC discussed the use of accommodations and the need for further study in this area, possibility funded by another grant.

Validation Process

The scope of work for the MWAC grant did not include validation processes, only the development and production of the assessment and associated materials. As the grant timeline came to a close, recommendations were made to the consortium states regarding what their next steps should be in validating the assessment within their state. Measured Progress recommended:

- aligning the MWAC items to the various state standards,
- validating the assessment in each state by assessing each state's own ELL population, and
- conducting standard setting to establish proficiency levels, once the assessment was used operationally

Additionally, MWAC applied for a renewal embedded assessment grant to assist the states in these next steps, but was not awarded such a grant.

Test Administration and Technical Manual

The Mountain West Assessment was not administered operationally during the grant timeline. For more information on test administrations, please see the *Pilot and Field Test* section of this chapter.

A technical manual for the Mountain West Assessment was not a required deliverable for the grant. In the place of a technical manual, a Final Report was written that included a summary of the work completed, changes from the original scope of work, and technical information on the pilot and field test administrations.

Scoring and Reporting

Measured Progress and the consortium were responsible for the scoring of the pilot and field tests during the grant. All multiple-choice items were scored electronically. All open-response items were scored onsite at Measured Progress by trained scorers using the same scoring and quality control procedures used for all the company's large-scale assessments. Each scorer received thorough training before scoring MWAC items and had their work monitored by quality assurance staff. Items for reading and writing were scored using score guides and rubrics developed by

RETURN to TABLE of CONTENTS

Measured Progress assessment and scoring specialists. This process was revised from the proposed scope of work, which stated that educators from each state were to be involved in the scoring of open response items. It was later decided that Measured Progress' trained scorers should score all field test items for consistency and accuracy. Of course, subsequent use of the materials by the states could make use of in-state scorers.

In hopes of better preparing the consortium states for the scoring of open-response items, a scoring institute was conducted. In late October 2004, two participants from each state were invited to the Measured Progress offices in New Hampshire to take part in a hands-on scoring institute. The purpose of the institute was to introduce participants to the processes and procedures used by Measured Progress when scoring open-response items, as well as give them the opportunity to use the Measured Progress computerized scoring system to score. The overall goal of the institute was to give states the information needed to better understand various pieces that create an accurate and secure scoring system. At the time of the institute, states were planning on scoring items in-state or with the support of a contractor.

Agenda topics at the Scoring Institute included:

- Overview of the Measured Progress scoring system
- Who and what kind of people do the scoring
- Training and qualification of scorers
- Security of test materials and student information
- Organization during the scoring process
- Preparation for scoring
- Scoring guides
- Benchmarking
- Control measures to prevent errant scoring
- Monitoring the scoring process

Institute attendees also participated in mock training and scoring sessions for reading and writing open-response items. After learning about specifics for scoring reading and writing items, reviewing score guides and rubrics, and training items, participants were given the opportunity to score MWAC open-response items using the Measured Progress system. In evaluations collected after the institute, participants noted a more in-depth understanding of the scoring process.

During the scoring institute, a production team was on hand to videotape the entire session. The video was then adapted with additional interviews and information into DVD format and distributed to each member state. Given each state's own plan for scoring, the DVD was designed to act as an introduction to the scoring of open-response items.

Conclusion

The Mountain West Assessment instrument created by the consortium is a promising new strategy for measuring ELL's English language proficiency for a variety of reasons. First, the assessment series assesses the English language skills required for the academic context of the mainstream American classroom. Second, the assessment provides a link between state English language development and academic content standards. Finally, the assessment can be used in conjunction with other measures of English language proficiency and student performance on other state assessments to get a more accurate picture of ELL's skills and readiness for achieving success in mainstream academic contexts.

Since the MWAC project, some of the participating states have created their own ELD standards and are using the MWAC instruments in a variety of ways. Some are using them as a source of items and some are using the tests on an interim basis until they create or adopt tests aligned with their unique ELD standards.

References

Abeela, R., Urrutia, J., Shneyderman, A. (2005). An examination of the validity of English language achievement test scores with an English language learner population. *Bilingual Research Journal*, *29(1)*, 127-144.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, England: Multilingual Matters.

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

Zehr, M.A. (2006). Team-teaching helps close language gap. *Education Week 26(14)*, 26-29.

RETURN to TABLE of CONTENTS

Chapter 4

The English Language Development Assessment (ELDA)

Julia Lara, Steve Ferrara, Mathina Calliope, Diana Sewell, Phoebe Winter, Rebecca Kopriva, Michael Bunch, and Kevin Joldersma

The No Child Left Behind Act of 2001 (NCLB; 2002) requires all states to assess the English proficiency of English language learners each school year. Under NCLB Title I and Title III, states are required to measure the annual growth of students' English language development in reading, listening, writing, and speaking, and comprehension toward attainment of full English proficiency. The English Language Development Assessment (ELDA) was designed to assess the development of proficiency in relation to the English language proficiency (ELP) standards of participating states.

The development of ELDA began in 2002 with efforts to assist states in meeting NCLB requirements. The Council of Chief State School Officers (CCSSO), along with states in the State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS), solicited proposals from test development organizations to work collaboratively with LEP-SCASS and CCSSO on ELP assessment development.[1] They selected the American Institutes for Research (AIR).

LEP-SCASS received an Enhanced Assessment Grant under Title VI (Section 6112) of No Child Left Behind (P.L. 107–110; NCLB) from the U.S. Department of Education to fund development, validation, and implementation of an ELP assessment. During fall 2002 through December 2005, LEP-SCASS, CCSSO, AIR, and Measurement Incorporated (MI) worked together to develop the ELDA. Nevada, as the lead state in the grant, and CCSSO managed the ELDA project for LEP-SCASS. The Initial Steering Committee included state education agency officials, assessment experts, linguists and ELL experts. The project included outside consultants to evaluate the development process and provide design and technical advice: the Center for the Study of Assessment Validity and Evaluation (C-SAVE) at the University of Maryland (validation studies) and

[1] The CCSSO/LEP-SCASS projects are networks of state education agency staff that combine their resources for the purpose of development of assessment-related tools and products that benefit the member states. There are 13 such state networks coordinated by CCSSO. Member states pay a yearly fee to the Council to defray the cost of travel, overnight accommodations, consultants, and administration. The LEP-SCASS consortium was formed at the request of state education agency officials interested in developing procedures, products and services focused on ELL students. Since its inception, the LEP-SCASS staff and consultants have produced guides for scoring ELL student responses to math and science items, a handbook for assessing ELL students, and research papers on ELL assessment issues.

RETURN to TABLE of CONTENTS

Measurement Inc., (K–2 test development, administration, scoring and reporting). At the start of the ELDA development project, 18 states were members of the LEP-SCASS. Thirteen states participated in the process of developing, field testing, validating, and implementing ELDA as an operational assessment.

The consortia determined that a valid ELP assessment for English learners in kindergarten through grade 2 (K–2) should rely upon observational data of English learners in natural classroom settings. For this reason, a separate test blueprint was developed for ELDA grades K–2 and ELDA grades 3–12 forms. It should be noted that the K–2 and the 3–12 versions of ELDA are both driven by theories of academic language and are both aligned to participating states' ELP standards. The test development process for ELDA grades 3–12 and ELDA grades K–2 are described in separate sections below.

ELDA Grades 3–12

The Theoretical Basis of ELDA

ELDA has been designed to assess the construct of "academic English" (Butler et al, 2004). The driving force—and the departure of this assessment from many ELP assessments—is the NCLB requirement that students classified as English-language learners be assessed annually in their progress towards proficiency in academic English. For purposes of test design and development, we defined academic English as falling into one of two categories: (1) language used to convey curriculum-based, academic content, and (2) the language of the social environment of a school. The concept of academic English is evolving, and it is important to make the point that although the ELDA items and prompts are written in the language of the classroom and of the academic subjects listed below, items do not require skills in or knowledge of content in those subjects. The concepts are not being assessed; the students' understanding of spoken and written texts about the concepts and their ability to write and speak about the concepts are being assessed. Any content a student is expected to use is provided in the stimuli or item prompt.

Within ELDA's four language skill domains—listening, speaking, reading, and writing—several academic content areas constitute the context for test items:

- English language arts
- Math, science, and technology
- Social studies
- School environment

This assessment is informed by second-language development theory of communicative competence which posits that ELP tests should measure communicative and participatory language in the context of the classroom and that they should be age/grade appropriate. This test is a departure from existing ELP tests in that ELDA measures English language mastery along the language development continuum and for each language skill domain. In addition, ELDA attempts to measure mastery of "academic English." Previous ELP assessments were designed to assist local educators with student placement decisions and measure low-level skills. Consequently, students were exited to the English-only classroom before mastery of the English language skills. Once in the mainstream classroom, ELL students were unable to meet the linguistic demands there and were often labeled as poor performers. Thus, limited proficiency in English was confounded with poor knowledge of the subject matter being taught. Moreover, the previous ELP tests were not able to provide instructionally relevant information, nor were they aligned to states' English language development standards.

Standards Used as a Base for Test Development

The starting point for the ELDA design was a synthesis of all state-level ELP standards that existed among the project's participating states. Of 18 states that initially formed LEP-SCASS membership, one-third had existing state ELP standards in each of the four domains of listening, speaking, reading, and writing.

The initial state ELP standards were carefully reviewed and merged by AIR staff. Then, project steering committee members agreed on a common core of standards for each domain by discussing standards they considered important and appropriate for ELLs in all LEP-SCASS states. They also considered which standards were appropriate at each grade cluster. Some states used ELDA's ELP standards to guide the development, revision, analysis, and adoption of their own ELP standards. Other states used them to review their existing ELP standards and ensure alignment with ELDA's.

State academic content and achievement standards are mandated by the U.S. Department of Education under NCLB for three content areas: reading/language arts, mathematics, and science. With reference to testing ELL students under Title III of NCLB, the law requires that English language proficiency standards be aligned with challenging state academic content standards and student academic achievement standards as described in Title I.

In order to align the test to the standards, CCSSO and AIR led a detailed and stakeholder-approved process of identifying ELP standards that could be used in test design, creating benchmarks from the standards, and developing items from these standards and benchmarks. In the case of academic content standards, the relationship to the test is less direct. Alignment between content standards and an ELP assessment is not implied in the non-regulatory guidance put forth by the U.S. Department of Education.[2]

Test Blueprint and Item Development [3]

AIR test developers and psychometricians drafted test blueprints and item specifications for each domain and grade cluster. To develop items that measure ELDA's ELP standards as specified by the content specifications, AIR brought together a pool of item writers which included external item writers, NAEP foreign language item writers, AIR content experts, and teachers with experience in test development who were recommended by LEP-SCASS states.

Listening Tests. ELDA Listening for each of the three grade clusters (3–5, 6–8, and 9–12) was designed to be administered through a cassette tape or compact disc (CD) medium. All test items are in the four-option multiple-choice format. Listening texts impart information drawn from the four topic areas: English/language arts; mathematics, science, and technology; social studies; and school environment. Text topics within the academic content areas were selected to avoid those that would typically be found in a grade-appropriate curriculum to ensure that the assessment would measure comprehension, not prior content-area knowledge. However, the texts reflect the discourse features typical of academic content areas. The operational forms for grade clusters 3–5 and 6–8 contain a total of 50 test items each. For grade cluster 9–12 there are a total of 60 multiple-choice items. High test-item totals for operational forms are a product of a five-level scale of performance.

Speaking Tests. ELDA Speaking is designed to be administered through a cassette tape or CD medium, thus eliminating written discourse from the measurement of an oral-based construct. It also can be administered orally to individual students. A test booklet containing graphics provides the student with some visual context for the audio prompts. The graphics are designed to help the student structure a response. Student responses to the prompts are captured on an individual student cassette recorder for off-site scoring. In operational administrations of the speaking test, schools may opt for oral administrations with local scoring, although the test content remains the same. Responses are scored on a 0–2 rubric. This rubric identifies:

- rhetorical features (i.e., organization of ideas and information, use of discourse markers to support organization),
- appropriateness (i.e., relevance and completeness),
- quality and quantity of the response (i.e., development and specificity, adequacy of the response in addressing the task), and
- correctness of spoken responses (i.e., appropriate vocabulary and comprehensible pronunciation).

The rubric and benchmark responses also account for consideration of audience.

Reading Tests. ELDA Reading for each of the three grade clusters has many of the design features of the listening test: four-option multiple-choice test format, identical operational test item numbers in each cluster, identical approach to topic content selection of reading texts, and similar distributions across the four topic areas. Each reading test is composed of three sections: Early Reading; Reading Instructions; and Reading Narrative, Descriptive, Expository, and Persuasive Texts. As with the listening forms, the reading operational forms for grade clusters 3–5 and 6–8 contain a total of 50 test items each. For grade cluster 9–12, there are a total of 60 multiple-choice items. High test item totals for operational forms are a product of a five-level scale of performance.

Writing Tests. ELDA Writing for grade clusters 3–5, 6–8, and 9–12 share some features with the listening and reading components—distribution across topic areas and an emphasis on the language of the classroom—with a logical difference: they also contains constructed-response items. Three main sections comprise the writing tests: multiple-choice editing and revising items (6–9 items per

[2] Final Non Regulatory Guidance on the Title III State Formula Grant Program Standards, Assessment, Accountability. Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students, February, 2003.

[3] See standards and specification document at: www.ccsso.org/projects/elda/Research_Studies.

RETURN to TABLE of CONTENTS

form), multiple-choice planning and organizing items (6 items per form) and a combination of short and extended constructed-response essay prompts (4–5 per form).

The editing and revising items are built around short stimuli, designed to simulate student writing. In some of these multiple-choice items, relevant portions of the text (a word or phrase), which may be grammatically incorrect, are underlined; students are asked to choose from options to replace the underlined text or to indicate that it already is correct. In other items, students are asked to choose an appropriate topic or concluding sentence or to provide missing information. The revising and editing items are designed to test students' ability to identify and correct sentence-level as well as text-level problems.

Pilot and Field Testing of Assessment Items and Tasks

In May 2003, 31 schools in 12 states participated in a pilot test of ELDA. The purpose of the pilot test was to determine whether test administration directions were clear for teachers and students, test administration procedures were feasible and efficient, and English language learners responded reasonably to the various item types. The pilot test included the reading, listening, writing, and speaking assessments for the three grade clusters. Schools were identified and recruited for participation so that the sample of schools was diverse in terms of type, size, location, and student demographics. Each school provided 10 students: five English language learners with low-to-intermediate English proficiency and five additional students that included a mix of English language learners with intermediate-to-high English proficiency, former limited English proficient students, and native English speakers. Participating students were drawn from each of grades 3–12 and reflected a diverse mix of race/ethnicity, sex, native language, country of birth, time in the U.S., and time learning English.

Pilot test students came from more than 20 language backgrounds and more than 30 countries. Results from item analyses, student focus group reports, and teacher reports indicate that students understood test administration procedures and were able to give their best performances in all four language skill domains. Test score reliabilities ranged from 0.77 to 0.92, similar to score reliabilities achieved in state content area assessments. Based on input from pilot test teachers, AIR and LEP-SCASS members revised administration procedures to make administering the test easier for teachers and taking the test easier for students. Analysis of results informed further item development.

A multi-state field test was conducted in spring 2004. The purposes of the 2004 field test were to (a) gather adequate data (i.e., 1,000 responses per item), evaluate items, and create the ELDA score scales; (b) assemble operational form 1 of ELDA Listening, Reading, Writing, and Speaking sections for use in 2005; and (c) conduct special studies relevant to the validity of interpretations about students' English proficiency from the ELDA scores.

Both field test and operational administrations were conducted in spring 2005. A field test was conducted in five states: Georgia, Indiana, Kentucky, New Jersey, and Oklahoma. The primary purpose of the 2005 field tests was to yield data on items to assemble operational forms 2 and 3 of assessments for use beyond 2005. Six states administered operational form 1 of ELDA and reported results to meet No Child Left Behind requirements: Iowa, Louisiana, Nebraska, Ohio, South Carolina, and West Virginia.

Validation: Psychometric Analyses[4]

Findings of classical item statistics which indicate the overall difficulty of the ELDA items and tasks suggest that the listening and speaking assessments were relatively easy for students in the field test (i.e., item difficulties in the range 0.70 to 0.81) and the item difficulties for the reading and writing items and tasks are in more typical ranges (i.e., 0.54 to 0.67). Items and tasks in the reading, listening, and writing assessments are moderate to strong (item-total correlations in the range 0.47 to 0.62 range) and strong for the speaking assessment (i.e., item-total correlations range from 0.81 to 0.87). (See Table 1).

Rates at which examinees do not respond to test items also are relevant to the difficulty of the items and provide some indication of the level of motivation that examinees displayed on the 2004 ELDA field test. Results indicate that students omitted few items in reading, listening, and writing. These rates compare favorably to non-response rates of native English speakers in academic content area assessments. The non-response rates are particularly low in the writing assessment, which contains short and extended constructed-response items which may be omitted by as

[4] For full report go to: www.ccsso.org/projects/elda/research-studies.

RETURN to TABLE of CONTENTS

Table 1. Mean Item Difficulty and Discrimination Statistics

	Item Difficulty	Item Discrimination
Reading	.61–.67	.56–.60
Listening	.70–.72	.60–.62
Speaking	.77–.81	.81–.87
Writing	.54–.59	.47–.53

Note. Ranges of means across forms and grade clusters in the 2005 field-test forms.

many as 5% of examinees in academic content assessments. The non-response rates in speaking are high, particularly in the grades 6–8 cluster. These rates may suggest a range of concerns about assessing English proficiency of English language learners (e.g., reticence in assessment situations), the delivery system for ELDA Speaking (i.e., prerecorded tasks delivered via audio recording; examinees record responses for subsequent scoring), the design of speaking tasks (e.g., the scaffolded prompts), or the difficulty and appropriateness of the tasks themselves (e.g., the degree to which the tasks offer opportunity for response for the diversity of English language learners who participated in the field test). (See Table 2).

Differential item functioning (DIF) indicates whether items function differently for examinees of equal proficiency from different subgroups. If items are unequal in difficulty for equally proficient members of different subgroups, the items function differently for the subgroups. This difference is considered unfair to the subgroup that finds an item more difficult. DIF is relevant to how valid any inferences are about an examinee's English proficiency based on his or her test performance.

Results show that relatively few items were flagged for DIF in all ELDA domains and grade clusters except in reading and listening grades 6–8 and speaking grades 9–12. LEP-SCASS reviewed all flagged items, suspended from subsequent use a small number of flagged items, and approved all other flagged items for subsequent use on operational test forms because they could find no content of contextual topics or features in the flagged items to explain the DIF flags and warrant discontinuing their use. (See Table 3).

Table 2. Non-Response Rates (in Percentages)

	Grade Cluster		
	3–5	6–8	9–12
Reading	1.8	1.3	2.3
Listening	0.3	0.5	4.3
Speaking	4.7	12.0	7.1
Writing	0.6	1.2	1.2

Note. Across 2005 field-test forms and grade clusters. Combination of items skipped and items not reached.

Table 3. Items Flagged for Differential Item Functioning (DIF)

	Grade Clusters		
	3–5	6–8	9–12
Reading	9/162 (6)	25/168 (15)	18/192 (9)
Listening	7/150 (5)	15/150 (10)	12/180 (7)
Speaking	3/60 (5)	0/60 (0)	18/60 (30)
Writing	4/76 (5)	3/76 (4)	1/80 (1)

Note. Across 2005 field-test forms within grade cluster. Comparisons are made for males vs. females, speakers of Spanish vs. other foreign languages, and students currently in limited English proficiency programs vs. students exited from such programs. Numbers of items flagged/total number of items; percentages in parentheses.

The degree to which the ELDA forms yield scores that are free of error are indicated using an internal consistency reliability estimate, coefficient alpha. The reliability estimates for all ELDA field-test forms exceed 0.85, except for ELDA Writing. The writing assessments are relatively short (i.e., 19 items for 28 points in the assessments for grades 3–5 and 6–8; 20 items for 31 points in the assessment for grades 9–12) and contain a variety of items types—multiple-choice (MC) and short and extended constructed-response (CR) items—that assess a range of writing skills (e.g., writing a draft, editing). These features explain the relatively low internal consistency reliability estimates for the ELDA Writing. (See Table 4).

AIR used Masters' Partial Credit Model (1982), an extension of the one parameter Rasch model that allows for both multiple-choice and constructed-response items, and widely used Winsteps software to estimate ELDA item parameters. Because each part of ELDA (i.e., the listening, reading, writing, and speaking domains) contains a

RETURN to TABLE of CONTENTS

Table 4. ELDA Score Reliabilities: Coefficient Alpha

	Grade Cluster		
	3–5	6–8	9–12
Reading	.93	.93–.94	.94–.95
Listening	.91–.92	.92–.93	.94–.95
Speaking	.88–.90	.93–.94	.88–.92
Writing	.76–.82	.84–.85	.84–.87

Note. Across 2005 field-test forms within grade cluster.

common set of items between adjacent grade clusters, the grades 3–5, 6–8, and 9–12 forms in each ELDA domain were jointly calibrated in a single Winsteps run for each subject. The joint calibration produced a common, vertically linked scale across grade clusters for each content area. For each Winsteps run, the mean of the item difficulty parameters was fixed to zero so that operational form 1 had an average difficulty (i.e., average item step value) equal to 0.0. (See Table 5).

We examined items that Winsteps flags for misfit to the partial-credit/one-parameter model. Misfit statistics indicate items that assess language and other proficiencies that may be related but tangential to the ELDA target construct, proficiency in reading, listening, writing, or speaking. Results indicate that 1 to 24% of items were flagged for misfit. LEP-SCASS reviewed all items flagged for misfit, suspended from subsequent use a small number of flagged items, and approved all other flagged items for subsequent use on operational test forms because they could find no features in the flagged items to explain the misfit flags and warrant discontinuing their use.

Table 5. Items Flagged for Misfit in IRT Calibrations

	Grade Cluster		
	3–5	6–8	9–12
Reading	33/162 (20)	27/168 (16)	32/192 (17)
Listening	36/150 (24)	34/150 (23)	35/180 (19)
Speaking	8/60 (13)	12/60 (20)	11/60 (18)
Writing	1/76 (1)	13/76 (17)	3/80 (4)

Note. Across 2005 field-test forms within grade cluster. Numbers of items flagged/total number of items; percentages in parentheses.

Validation: Validity Studies[5]

CCSSO's LEP-SCASS technical advisory committee, the Center for the Study of Assessment Validity and Evaluation (C-SAVE), and AIR developed a validity research agenda. C-SAVE performed two general types of analyses, item-level and test-level analyses, with several analyses conducted in each category to provide forms of evidence.[6]

The purpose of the item analyses was to assist CCSSO and AIR with selecting items from the pool of field-tested items to form valid ELDA forms. These analyses supplemented AIR's traditional item analyses that focused on scoring keys and rubrics, item difficulty assessments, biserial and point biserial discrimination indices, and DIF. [7]

Latent Class Analyses. The main purpose of the ELDA field test was to evaluate the initial pool of test items. We constructed a different collection of items for each domain (reading, writing, speaking, and listening) in each grade cluster (3–5, 6–8, and 9–12). Each collection was assigned to one of two field-test forms (A, B) so that each form reflected, as closely as possible, the final test blueprint. The field-test data set included item responses for every item for each student together with collateral data (e.g., language acquisition level, primary language, type of English for Speakers of Other Languages (ESOL) program, standard demographics) on every student. Using the Winmira program, we fit five-class models to item response data from each field-test form.

For each domain and grade-cluster form, we estimated the proportion correct for each item within each latent class. To evaluate the validity of items for discriminating among the ordered latent classes, we calculated the differences in proportion correct between adjacent classes (See Table 6).

Teacher Ratings of Student Proficiency. The field-test data collection included teacher assessment of each student's language proficiency in reading, writing, speaking, and listening. These took the form of a 5-point devel-

[5] This section is based on and contains excerpts (with permission) from Kopriva, R. (October, 2004). Field Test *Validity Study Results: English Language Development Assessment. Final Report.*

[6] C-Save Center was formerly of the University of Maryland and now is housed at the University of Wisconsin.

[7] See the full validity report in: http://www.ccsso.org/projects/elda/Research_Studies.

RETURN to TABLE of CONTENTS

opmental scale in each language skill domain. For each domain, these data were used to group students by level and calculate the proportion correct on each item in every form for each proficiency level. As one would expect, the proportion correct increases with proficiency level. In order to evaluate the validity of items for discriminating student proficiency levels as reported by the teachers, the differences in proportion correct between levels were calculated as was done for the latent class analysis results (See Table 7).

Table 6. Latent Class Analysis

Item Order	Class A	Class B	Class C	Class D	Class E
1	0.51	0.87	0.96	0.98	1.00
2	0.62	0.98	1.00	0.98	1.00
3	0.63	0.98	1.00	0.99	1.00
4	0.45	0.78	0.88	0.93	0.97
5	0.35	0.61	0.85	0.95	0.97
6	0.22	0.48	0.76	0.83	0.93
7	0.39	0.86	0.97	0.98	0.99
8	0.25	0.73	0.94	0.97	0.99
9	0.41	0.62	0.87	0.94	0.98
10	0.51	0.84	0.98	0.99	0.98
11	0.28	0.65	0.91	0.98	0.99
12	0.26	0.64	0.93	0.96	0.98
13	0.16	0.20	0.33	0.50	0.61

Developmental Level Ratings of Items. To analyze the developmental level of items, experts with extensive expertise in ESOL instruction and language testing were trained and charged with assigning to each item the performance level designation (i.e., beginning, lower intermediate, etc.) that best identified the level of English language development at which the item was focused.

Item Analysis. In addition to rating item developmental level, we used two other sources of information as criteria for judging the items: the latent class gradient results and the student proficiency gradient results. In order to evaluate the items consistently over domains, grade spans, forms, and item types, a flagging system was developed that would identify the strength or weakness of each item as referenced to a set of criteria. The criteria were based on the degree to which the item discriminated at a single location on the developmental scale and the consistency of evidence across the three sources. The results of the item reviews according to these criteria were used—along with the traditional item analyses produced by AIR—to determine whether each item should be considered for the operational forms of the ELDA or revised or discarded.

Table 7. Proportion Correct by Student Proficiency Level

	Student Proficiency Ratings (PR)				
Item Order	1	2	3	4	5
1	0.55	0.87	0.95	0.98	1.00
2	0.68	0.96	0.98	0.99	1.00
3	0.68	0.97	0.98	1.00	0.99
4	0.48	0.82	0.89	0.91	0.93
5	0.41	0.67	0.86	0.89	0.97
6	0.25	0.59	0.74	0.78	0.90
7	0.48	0.87	0.94	0.97	1.00
8	0.42	0.79	0.90	0.92	0.97
9	0.47	0.68	0.88	0.92	0.94
10	0.64	0.86	0.95	0.96	0.99
11	0.36	0.77	0.87	0.93	0.97
12	0.38	0.72	0.88	0.93	0.96
13	0.18	0.25	0.36	0.43	0.58

Results of Item Analyses. The developmental level ratings of the items found that, with the exception of speaking, all domains contained field-test items representing each of the five developmental levels in each grade cluster, with fewer items, in general, at levels 1 and 5. The speaking domain had no items that were rated level 5. The reading items in grade cluster 9–12 received "strong" flags, indicating potentially weak items, more often than those in grade clusters 3–5 or 6–8. The flagging pattern for listening items was more consistent across grades. A strong flag suggests that the item discriminates poorly or that all three sources conflict on where the item discriminates. The writing domain's multiple-choice items received the highest proportion of "strong" flags overall, and only a few speaking items in the 9–12 cluster received "strong" flags.

Analyses of Relationships among Development Level of Items, Percent Correct and Teacher Ratings of Student

In addition to analyzing individual item validity, we evaluated the relationship of developmental level ratings of items and teacher ratings of student proficiency to item difficulty. This was done by developing cross tabulations of percentage correct by item developmental level ratings and student proficiency ratings, and by performing two-way

RETURN to TABLE of CONTENTS

mixed-model ANOVAs to estimate how difficulties varied over the two factors.

The results of the ANOVAs, using percent correct as the outcome variable, were remarkably uniform over domains and grade clusters. In all cases, the main effects (development level ratings and teacher rating) were significant. The interaction was also significant for almost all domains and grade clusters, except for Reading 9–12A and B, Listening 3–5A and 6–8A, and Writing (MC) 9–12A and B. Although the findings were not significant, the results were not disordinal and continued to reflect the monotonic nature of the other analyses.

Across item developmental levels and across student proficiency levels (and for all domains and grade spans), item probabilities were ordered monotonically. That is, for items in developmental level category 1, the percentage correct uniformly increases by student proficiency level. Likewise, for student proficiency level 1, the probability correct decreases as items become more difficult in the higher developmental levels. This occurs over all developmental levels, for all domains, and in all grade spans. The results validated both teacher rating of student proficiency and the expert development level rating of items in terms of logical expectation about percents correct.

Analyses of Field Test Scores

We reviewed the quality of ELDA by how the field tests as a whole measured the targeted sets of latent traits inherent in the four language domains of reading, writing, speaking and listening. These analyses were not performed on final, operational test forms. The results can be generalized somewhat to the operational forms, because test construction was conducted using the findings of the item analyses, but the results do not apply directly to operational forms. Three sets of analyses were conducted:

- The relationships among ELDA, the Language Assessment Scales (LAS), the Idea Proficiency Test (IPT), and teacher ratings with respect to how well they interpret the four language domains.
- The underlying internal developmental structure of ELDA, specifically, the theoretical nature of the development of proficiency in one language for those whose primary language is another.
- The latent class framework of the language skill domain scores in terms of proficiency level and in terms of item group indicators that cut across performance levels.
- Other measures of proficiency in reference to the judgment valuations of the complexity of skills measured in ELDA items.

Relationship of ELDA with Other Measures

We investigated the relationships of ELDA scores to other measures of English language proficiency—LAS proficiency levels, IPT proficiency levels, and teacher ratings of student proficiency—for the students in the field test. We also investigated relationships for critical subgroups identified by LEP-SCASS: language proficiency level, including post-ESOL and native English-speaking students; language/linguistic group; type of ESOL instruction; and grade level. For each grade cluster, the resulting multitrait-multimethod matrix was represented as a path model. Four latent traits were included in this model to represent the true scores on the reading, writing, listening, and speaking traits. The other latent variables (LAS, IPT, and ELDA proficiency levels and teacher ratings) were included to represent the effects of the methods.

Overall, the findings suggest convergent validity of methods across language skill domains and some evidence of discriminant validity. ELDA, LAS, IPT, and teacher assessment all measure language proficiency, but in all clusters (especially at 6–8), there appeared to be only limited ability of the assessments to discriminate the language skill domains within the measurement of language development. This remains a substantive question: How much unique variance within each domain should one expect to build into an assessment of language proficiency?

We fit models for each group within the four subgroup categories. With very few exceptions, the ELDA and teacher rating score-trait correlations were higher than either LAS or IPT. For the most part, ELDA behaved very similarly to teacher ratings, while LAS and IPT loadings tended to be analogous. In general, ELDA loadings were respectable in size and stable across language skill domains tested, suggesting stability over most groups within each of the subgroups. They also clearly differentiated the ELDA findings from the other tests, which in turn tended to consistently produce substantially lower score-trait correlations.

Latent Class Analyses of Field Test Scores

To analyze the underlying internal developmental structure of ELDA, we performed a standard latent class analysis on the total test scores in each domain and the mixed Rasch latent class method of analysis. The framework for these latent class analyses is the theoretical view of English language development in which an English language learner passes through multiple stages of development, from pre-production to advanced fluency, in each of four major modes—listening, speaking, reading and writing—that are reflected in the four domains assessed by ELDA. These separate stages are interdependent in that, e.g., listening must be at least partially developed before speaking, reading, or writing can be initiated.

The standard latent class analyses performed on the field-test forms were generally consistent with an ordered five-class model that captured the developmental stages of the English language development process. In contrast, the mixed-Rasch latent class method resulted in five classes for approximately 60% of the domain/grade-cluster/form combinations. Lack of fit was particularly evident in the speaking domain, where no field-test form supported five classes, and in the writing domain, where only two forms supported five classes. It is unclear whether this finding results from the developmental process being less refined for writing and speaking than for reading and listening, whether the field-test forms for writing and speaking were not sufficiently valid to allow five-stage discrimination, or whether speaking and writing are multidimensional, as measured by the field-test forms.

Analyses of Developmental Level Structure

The complexity of skills assessed by the items was analyzed using a simplex structural model (defined below). The model is based on the expert judgment valuations of the complexity of skills required of the items, the developmental level ratings. In addition, the complexity of the structure of the items was evaluated relative to IPT and LAS proficiency level scores.

To prepare for these analyses, items were identified by developmental levels, as defined by complexity of skills required, by expert judges using ELDA field-test item results. Because the number of items in the highest and lowest groups was too small to generate a stable score, items were grouped into three categories of development: levels 1 and 2 were identified as low English proficiency, level 3 formed the medium developmental level, and items in levels 4 and 5 were assumed to be those that discriminated primarily at the high level of proficiency. Mean percentage correct scores in each of the three developmental levels were computed for each student and formed the basis of the analyses.

The data were assessed using a developmental model representing a simplex structure, via a structural equation model to fit recursive regressions for the various grade clusters and domains. The simplex structure assumption posits that skills for the most part are cumulative: more complex skills build on simpler skills for most language constructs. Overall, the results support this hypothesis over the grade clusters. Reading and listening models generally indicate a good-to-excellent fit, speaking and writing constructed-response models suggest an adequate-to-good fit, and writing multiple-choice findings indicate mixed but typically supportive models of fit. With some exceptions, the mean percentage correct monotonically decreased with complexity of skills being measured. Importantly, residual correlations between non-adjacent categories of complexity tended to be fairly low over grade clusters and domains, generally supporting the simplex notion.

Regressing Other Measures on Developmental Level Ratings

One of the primary purposes of the ELDA is to measure complex academic language proficiency skills in addition to the less complex skills addressed in more basic academic situations and in social language competency. Given the confirmation of this structure in the simplex analyses, the LAS and IPT language proficiency levels were regressed on the complexity of skills as defined in the expert judgment complexity valuation of the ELDA scores.

The regression analyses of LAS and IPT on these complexity categories found that ELDA and the other dependent measures were measuring considerably different skills, especially for writing, listening and speaking. In many cases, marginal amounts of information about skills in the LAS or IPT can be predicted from our understanding of complexity, as operationalized by ELDA. Most elusive overall is the ability of the LAS and IPT to predict higher level complexity skills. This finding is consistent with the evaluations of commercially available tests in the literature and by the ELDA development committees, where one of the main goals of ELDA was to measure a broader range of skills—particularly higher academic proficiency skills—than the tests that were currently on the market.

Assembling Operational Test Forms for Administration in 2005–2007 and Beyond

ELDA content specialists at AIR assembled draft versions of operational forms 1 (i.e., after the 2004 field-test analyses) and operational forms 2 and 3 (i.e., after the 2005 field-

RETURN to TABLE of CONTENTS

test analyses) in all content areas and grades. Each draft form underwent three levels of review by other ELDA content specialists, the ELDA development leader, and the ELDA project director. During each phase of review, these assessment specialists worked with AIR psychometricians to ensure that each form balanced the content and statistical requirements in the ELDA specifications. LEP-SCASS reviewed and approved operational forms 1, 2, and 3.

AIR content specialists assembled forms to meet the following specifications:

- The specified numbers of vertical linking items (i.e., common items across grade-cluster assessments) in operational form 1, horizontal linking items in operational forms 2 and 3 (i.e., common items across the same grade clusters in all three forms)
- Test blueprint features, such as the number of items per standard on each test form as a whole, the balance of benchmarks and content areas (e.g., mathematics, science, and technology; school/social environment), and the maintenance of item type order, as indicated in the ELDA specifications
- Miscellaneous features, such as multiple-choice item key counts and balance, passage topic balance, gender balance, item and classification soundness, and content overlap

Finally, items in all forms were sequenced for consistency with the relative position of linking items in operational form 1 and the field-test position of all other items being used operationally for the first time in operational forms 2 and 3. These requirements were reviewed for each form and across forms in grade clusters and content areas to ensure that all assembled forms were as parallel as possible from a content and statistical perspective.

Standard-Setting Process

At the initial meeting of the steering committee, December 2002, member states in consultation with AIR staff, decided on five ELP performance levels. Extensive discussion took place regarding the objectives of the assessment to be developed, the breath of content coverage to be assessed, the linguistic demands of the content area under consideration, and the number of items needed in order to cover the standards, and lastly, the amount of time it would take to complete each of the domains tested. As the development evolved, AIR and steering committee members further refined the performance level descriptors that were instrumental in the development of the performance standards.

Performance standards for ELDA grades 3–12 were set in August 2005. For the reading, writing, and listening domains, MI used a bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001). In this procedure, standard setters evaluated specially formatted test booklets and placed bookmarks at points where the difficulty of items appeared to change in ways that differentiated between adjacent performance levels (i.e., *pre-functional, beginning, intermediate, advanced,* and *fully English proficient*). For the speaking test, MI used a generalized holistic approach (Cizek & Bunch, 2007). In this procedure, standard setters evaluated live samples of student work, placing them into one of the five categories (*pre-functional to fully English proficient*).

Standard setters worked in grade-cluster groups (3–5, 6–8, and 9–12) to set standards for all tests in reading, writing, and listening. A single group set standards for all speaking tests. At the close of a three-day, standard-setting meeting, the individual groups turned their recommendations over to an Articulation Committee composed of representatives of each of the four initial groups. The function of the Articulation Committee was to merge the individual grade-cluster performance standards into a set of standards that would span the grades, eliminating or smoothing any cluster-to-cluster disparities or discontinuities they might find. The Articulation Committee also recommended procedures for combining scores to produce comprehension and composite scores. All cut scores were subject to final review and approval by CCSSO.

ELDA GRADES K–2

Theoretical Basis

The design of the ELDA for grades K–2 was informed by current views of early childhood development and implications for assessment. Between the ages of 5 and 8, children grow and change rapidly in terms of their motor, language, cognitive and social-emotional development. Consequently, in the development of the assessment, special attention was devoted to the overall time of the student observation, the format of the assessment, the interactions between teacher and student, the supports available to teachers and students (e.g., pictures, manipulatives), and the complexity of the language of the prompt. Given the dearth of empiri-

RETURN to TABLE of CONTENTS

cal data regarding best practices of assessing young English language learners, the member states and test developers relied on the research work that focuses on second language learning for the very young and the instructional practices appropriate for young English language learners. This effort was coupled with on-going input from expert consultants and teachers from member states.

Defining/Using State Content Standards

The process for incorporating state content standards into test specifications for ELDA grades K–2 was identical to that used for grades 3–12 (cf. American Institutes for Research, 2003). In November 2003, members of the ELDA K–2 advisory sub-committee met with staff from AIR to review state ELP content standards and select standards appropriate for ELDA grades K–2. Members and CCSSO staff combined identical or similar content standards from member states, eliminated those whose evaluation would be beyond the scope of the proposed methodology of the assessment, and prepared a final, consolidated set of content standards.

The condensed standards and ELDA K–2 framework were generated by AIR with support from early childhood education consultants and ELDA K–2 subcommittee members of the LEP-SCASS states. MI development staff reviewed the standards accepted by the membership in preparation for item development.

Test Blueprint and Item Development

Early in the process of developing ELDA for grades K–2, state members opted for an inventory approach over the traditional multiple-choice and constructed-response item approach because of the age and developmental stage of the student population. Each "item" in the initial inventories was a statement regarding an observable student behavior such as the following:

- Follows a two-step verbal instruction in a non-academic setting (e.g., going to the lunchroom)
- Identifies a picture of an object with the same ending sound as 'cat'
- Uses correct English words for manipulatives (content-, age-, and grade-appropriate items)

MI invited nine classroom teachers to participate in an item development session in Durham, NC, in February, 2005. The teachers, who were drawn from member states, worked with MI staff and the chair of the LEP-SCASS ELDA K–2 subcommittee to create inventory entries for ELDA grades K–2 Listening, Reading, Speaking, and Writing. The goal of this collaborative work was to generate enough teacher observations/items to construct three field-test inventories.

The item writers used the lists of standards and benchmarks collected from the LEP-SCASS member states in the consortium as their guides, along with other materials they brought to the session and those supplied by MI.

The final step in item development was the selection of anchor items from the current tests for grades 3–5. The anchor items were selected in order to link scores of ELDA grades K–2 assessments to those of the assessments for grades 3–12. These items were selected on the basis of their relevance to the grades K–2 content objectives and the fact that they were among the easiest of the ELDA grades 3–5 items, suggesting that they would not be too difficult for those students in grades 1–2.

Field Testing

Six states (Indiana, Kentucky, Nebraska, New Jersey, Oklahoma, and West Virginia) participated in the field test, with a total of 2,431 students (745 kindergarten, 831 grade 1, and 798 grade 2). MI scoring leaders conducted training for scorers as they did in 2004, and those scorers evaluated student responses to the writing prompts as well as responses to the speaking prompts.

Reliability and Validity

The ELDA K–2 inventories were administered in the spring of 2005 in their preliminary (long) version. Results are documented in CCSSO (2006) and summarized here.

Item face validity. MI staff, CCSSO staff, and two nationally recognized content experts met for a face-to-face review session in CCSSO's offices in Washington, DC, on March 11, 2005. This session was similar to those conducted in 2004 for items developed for grades 3–12. At the end of the review session, MI staff documented all recommendations, made the necessary modifications, and submitted all items to CCSSO for final approval. The basic structure of the inventories was validated by the two content experts, who provided suggestions for refocusing specific inventory entries (items) and approved the instruments for field testing.

Item reliability. Corrected item/total correlations for all inventories (items) ranged from 0.48 to 0.87, indicating an extremely high internal consistency as measured at the item level.

RETURN to TABLE of CONTENTS

Item response vs. teacher rating. The same analyses revealed correlations between item scores and teacher ratings ranging from 0.24 to 0.65, with most (60 out of 63) above 0.3, and 40 out of 63 had correlations above 0.5.

Item response by grade. The technical report includes analyses of item response by grade. With one exception, all grade-to-grade differences in item scores were positive (Speaking item 12 had a difference of 0 from K to grade 1). In general, differences between reading and writing were much higher (on average a full point from K to grade 2 in reading and just under a full point from K to grade 2 in writing) than between listening and speaking (about a quarter of a point from K to grade 1 for both and just under half a point from K to grade 2 in both). Overall, however, the indication is that all but one item show gains from grade to grade.

Test reliability. Generalizability analyses showed the inventories to have reliability coefficients ranging from 0.92 (listening, 7 inventories) to 0.97 (reading, 29 inventories). A reliability coefficient of 0.90 is considered to be excellent for individual decisions about students.

Test score vs. teacher rating. Correlations between inventory scores and teacher ratings of student proficiency ranged from 0.57 (listening) to 0.68 (reading and speaking). The correlation for writing was 0.58. All of these correlations reveal a strong relationship between scores on the inventories and classroom teacher judgments about students' levels of proficiency.

Operational Form Results

In the spring of 2006, ELDA grades K–2 (shortened version) was administered to 21,228 students in four states. Analyses were similar to those performed in 2005.

Item reliability. Corrected item/total correlations for all inventories (items) ranged from 0.58 to 0.86, slightly higher than in the field test, again indicating an extremely high internal consistency as measured at the item level.

Item response vs. teacher rating. The same analyses revealed correlations between item scores and teacher ratings ranging from 0.45 to 0.77.

Item response by grade. For the operational forms, the entries for the kindergarten level are different than those for grades 1–2; therefore, direct comparisons are available only for grades 1–2. All differences were positive, ranging from 0.10 (reading, item 1) to 0.47 (reading, item 4), with a mean of about one-fifth of a point from grade 1 to grade 2 for a given item.

Test reliability. Even though all inventories except listening were considerably shorter than the prior version, (reading, for example, changed from 29 inventories to 14), all reliability coefficients were above 0.90. Test reliability ranged from 0.94 (listening, all grades) to 0.97 (reading, grade 1). Given this range, it is safe to conclude that there is little variability at all in total test reliability, and that the predictions based on the field-test results are quite accurate.

Test score vs. teacher rating. Correlations between inventory scores and teacher ratings of student proficiency ranged from 0.65 (writing for kindergarten) to 0.77 (reading for grades 2 and 3). All correlations reveal a strong relationship between scores on the inventories and classroom teacher judgments about students' levels of proficiency.

Standard-Setting Process

Performance standards for ELDA grades K–2 were based on performance level descriptors developed specifically for these assessments by Malagon, Rosenberg, & Winter (2005). For ELDA K–2, the holistic approach was used for all inventories. Performance standards were set in a web conference in January 2006 and confirmed at a second web conference in July 2006. Details of conferences, procedures, and outcomes are described in Bunch & Joldersma (2006).

Creating Operational Forms

Subsequent to the 2005 field test, there were two key meetings concerning ELDA grades K–2. The first was in Savannah, Georgia, on July 6–7, 2005. At this meeting, state representatives presented many of the concerns voiced by K–2 teachers regarding the length and difficulty of administering the inventories. In August, an early childhood ELL expert joined the MI team of developers to begin revising the inventories with two key goals: shorten the inventories and provide more support for teachers who administer them.[8] On December 7–9, 2005, member state representatives met again in Washington, DC, to review revised materials. These materials were ultimately approved with modifications during December 2005 and January 2006. The final materials were submitted and approved on January 31, 2006.

[8] Dina Castro, from the Frank Porter Graham Child Development Institute, reviewed the inventories to assess developmental and linguistic appropriateness.

RETURN to TABLE of CONTENTS

Accommodations and Validity (K–12)

The administration manual of ELDA sets forth guidelines for offering and using accommodations for ELL students with disabilities. The recommendations were informed by members' understanding of special education requirements and extensive consultations with knowledgeable experts. Generally, the guidelines recommend that accommodations should always be related to a student's specific disability, and that they be consistent with those allowed in a student's IEP or 504 plan and with practices routinely used in a student's instruction and assessment. Since ELDA is a language assessment, and most accommodations offered to ELL students are language related, only certain types of accommodations are recommended with ELDA: computerized assessment; dictation of responses; extended/adjusted time; and individual/small group administration. There was recognition among member states that the research evidence for the use of these accommodations is limited and that more research is needed on the validity of such accommodations. Finally, in addition to those listed above, Braille and large print versions of the ELDA Reading and Writing are permitted. The listening and speaking tests were not produced in large print and Braille formats because the students respond to audible stimuli for these tests.

In 2005 and 2006, the test developers produced large-print and Braille versions of ELDA grades 3–12 Reading and Writing and shipped them as requested by schools. In addition to modified test booklets, other permissible accommodations for the ELDA administration included computerized assessment (typing of open-ended writing responses), extended/adjusted time for completion of the assessment, individual/small group administration, dictation of responses for all parts of the assessment with the exception of the constructed-response writing items, and any accommodations provided for under an individual student's documented IEP or 504 plan.

Test Administration and Technical Manual (K–12)

ELDA grades 3–12 was designed to be administered to class-sized groups of students simultaneously, with the exception of the ELDA Speaking. This assessment is scored live on site by teachers. In 2005 and 2006, separate ELDA Speaking Scoring Guides were developed containing text of the audible questions that students respond to orally, and sample answers to each question representing each score point. There was a scannable answer document for each student. During the test, the teachers filled in bubbles to score students' answers, and the scores were captured.

In contrast, ELDA for grades K–2 consisted entirely of inventory items completed by teachers. Teachers recorded the scores for each item in each student's test booklet, guided by information in the 2005 ELDA K–2 Test Administration Manual, and in 2006, by the ELDA Test Administration Manual and Teacher Support Materials. Because the inventories differ between kindergarten and grades 1–2, two versions of the Teacher Support Materials were developed.

At the conclusion of testing in 2005 and again in 2006, AIR and MI collaborated to produce technical manuals. These manuals (AIR, 2005; CCSSO, 2006) describe in detail test development, administration, scoring, analyses, and outcomes.

Scoring and Reporting (K–12)

Because K–2 inventory scores were recorded directly into student test booklets and not onto scannable forms, trained operators recorded the inventory scores in a tested data entry application. The ELDA Student Background Questionaire (ESBQ) for each student was then scanned, and the demographic data was captured and stored in databases divided by state. The scores were then merged with the demographic data via a unique matching bar code on each student's test booklet and ESBQ.

For ELDA grades K–2, MI staff key-entered identification information and inventory scores to a data file using a double-entry procedure. Entries were post-edited for out-of-range entries and other anomalies prior to uploading to score reporting programs. Score reporting was the same as for the ELDA grades 3–12 versions.

For grades 3–12, trained readers scored the constructed-response writing items according to initial range-finding results. CCSSO conducted a range-finding with participants from the ELDA consortium states in 2004. Supervised readers assigned scores based on the scoring protocols determined at range-finding, and then bubbled in their scores on scannable scoring monitors. Ten percent of all responses received a second reading to verify reliability of readers' scores. The monitors were then scanned, and the results were merged with the data derived from the students' multiple-choice answer selections.

RETURN to TABLE of CONTENTS

After all scores were processed, MI created the following PDF files for each district:

- Demographic Report
- District Summary Report
- Student Roster
- Individual Student Report
- Student-level Data File

The PDF files for each district were then transferred to a CD and sent to each district. Each state also received a CD with their corresponding files. The data sets included not only raw scores and levels, but scale scores, school and district information, student demographic data, and program information. Score reports included an interpretive section which explained the five performance levels, provided scale score ranges for all performance levels for all tests, and described how the comprehension and composite scores were derived.

References

American Institutes for Research (2005a). *English language development assessment test specifications and standards document*. Washington, DC: Author.

American Institutes for Research (2005b). *English language development assessment (ELDA) Technical Report: 2005 Operational and field-test administration*. Washington, DC: Author.

AIR/CCSSO/LEP-SCASS (undated). *English language development assessment K – 2 standards and benchmarks*. Washington, DC: AIR.

Brennan, R. L. (1983). *Elements of generalizability*. Iowa City, IA: American College Testing Program.

Bunch, M. B. (2006). *ELDA standard setting final report*. Durham, NC: Measurement Incorporated.

Bunch, M. B. & Joldersma, K. (2006). *Setting standards for ELDA K–2*. Durham, NC: Measurement Incorporated.

Butler, F. A., Lord, C., Steven, R., Borrego, M., & Bailey, A. L. (2004). *An approach to operationalizing language for language test development purposes: evidence from fifth-grade science and math*. (CSE Report 626.) Los Angeles, University of California: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Council of Chief State School Officers (2006). *English language development assessment K–2 Technical Manual: Spring 2006*. Washington, DC: Author.

Cummins, J. (1979) Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19*, 121–129.

Malagón , M H. , Rosenberg, M. B., & Winter, P. (2005). *Developing aligned performance level descriptors for the English language development assessment K–2 Inventories*. Washington, D.C.: CCSSO.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Editor), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum.

RETURN to TABLE of CONTENTS

Chapter 4

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

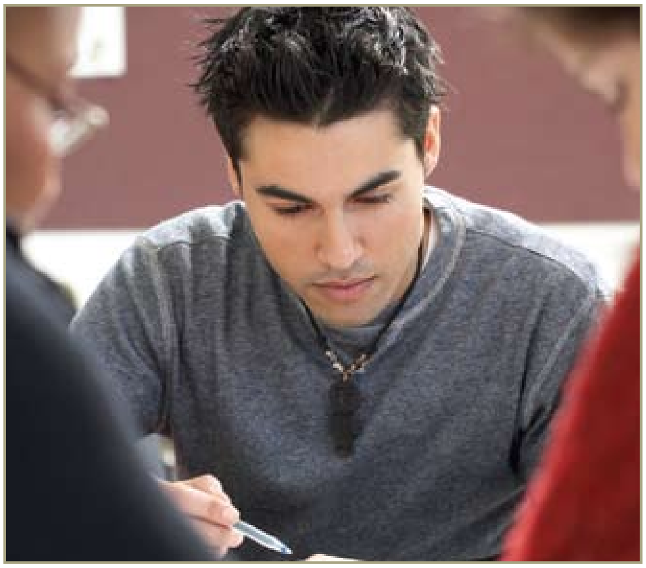
Chapter 5

Designing the Comprehensive English Language Learner Assessment (CELLA) for the Benefit of Users

Theodor Rebarber, Paul Rybinski, Maurice Hauck, Robin Scarcella, Alyssa Buteux, Joyce Wang, Christine Mills, and Yeonsuk Cho

Title III of the No Child Left Behind Act (NCLB; 2002) has had a profound impact on the English language proficiency assessment of students throughout the United States who are classified as English language learners (ELLs). By requiring that the language proficiency of all ELLs from kindergarten through grade 12 be assessed annually with a standards-based instrument that tests listening, speaking, reading, writing, and comprehension, Title III has ushered in a new generation of English language proficiency (ELP) tests.

The spur of NCLB, however, created an opportunity to go beyond federal requirements and to fundamentally rethink assessment for ELL students. Toward this end, AccountabilityWorks (AW), a nonprofit organization dedicated to fostering well-designed assessment and accountability systems, brought together a diverse but manageable number of reform-minded states sharing the same fundamental vision and eager to embrace this bold endeavor. The five consortium states and AW were joined by an outstanding test development partner, ETS, which contributed a team of talented, nationally respected designers and researchers. ETS products, including TOEFL® (Test of English as a Foreign Language™) and TOEIC® (Test of English for International Communication™), have been among the most widely used and most respected English language assessments in the world. In recent years, ETS has also developed English language proficiency assessments at the K–12 level, including work on assessments for New York, California, Puerto Rico, and the State of Qatar. With the support of a grant from the U.S. Department of Education, AW managed the work of the consortium while subcontracting much of the test development to ETS.



Consistent with the vision of the consortium, the Comprehensive English Language Learner Assessment (CELLA) was designed to meet four objectives that go beyond federal statutory requirements:

1. To reflect the nature of English language acquisition and its implications for assessment, including the extraordinary diversity of ELL students' English and native-language literacy skills at all ages (K–12) as well as the importance of growth measurement in determining program effectiveness (given that, by definition, ELLs are not proficient in English)
2. To base the assessment on the best scientific-based research available, including research on

RETURN to TABLE of CONTENTS

academic English, reading acquisition, and other design elements

3. To anticipate the needs of educators and other users at the local level by creating an assessment that is efficient to administer and provides data that is useful for improving instruction
4. To meet the most rigorous psychometric standards for validity and reliability—even in challenging areas such as speaking—and at the earliest grade levels

The importance of the consortium's assessment design goals are better understood by examining the ELL population, the process of becoming proficient in English, and the challenges of assessing both students and programs.

Theoretical Underpinnings of CELLA

ELL Students, the Nature of ELP, and Implications for Assessment

CELLA was designed to reflect the reality of English language instruction and acquisition in diverse schools facing today's challenges. While ELL students represent a culturally and linguistically diverse population, research permits us to describe such students as well as their skills. ELL students typically come from homes where another language is used. They may be bilingual, limited in their ability to use English, or monolingual (August & Hakuta, 1997). They may be born in the United States or recent arrivals to the United States—place of birth does not determine proficiency in English.[1] Many make rapid, continuous progress developing reading, writing, speaking and listening skills in English, while others make little progress in specific skill areas. Still others seem to reach a plateau in the development of skills, seemingly ceasing to acquire English altogether. Some seem to backslide, losing valuable language skills the longer they live in the United States. English learners represent a wide range of language backgrounds: over 80 different languages. The majority speak Spanish as a first language and come from such diverse countries as Argentina, Cuba, Mexico, Nicaragua, and Puerto Rico. They may be highly literate in their first language and proficient readers of English or, regardless of their age, incapable of reading in either their first or second language. They may be well educated in their home countries, or they may have enormous educational gaps when they begin their schooling in the U.S. They may have received excellent instruction in our U.S. schools or very poor instruction. Expressions, topics, genres, and situations familiar to some learners may be completely unfamiliar to others.

Grade levels do not determine the English proficiency of ELL students. ELL students entering kindergarten may have been born in the U.S. and may have picked up some English and literacy skills prior to their first day in school. On the other hand, some newly arrived ELL students may enter high school with no significant English skills and very low literacy in their native language. Nearly all variations between these two poles are not only possible; they are, in fact, present in the U.S. school system.

Every state defines the attainment of English proficiency differently. Because ELL students, by definition, are defined as those who have not reached adequate proficiency in English and because those who reach proficiency are exited from ELL status, every state defines the ELL population differently because the inclusion of students at the *near-proficient* to *proficient* levels vary.

A student who has exited an ELL program is no longer required to take a English language proficiency test. While a secondary use of CELLA is to help determine whether students are prepared to exit ESL or bilingual programs, the primary use, as required by NCLB, is the accountability of the programs providing students with such services. To this end, determining whether students have or have not attained proficiency in English is of limited use in assessing a program. Even measuring the time required to exit students is heavily influenced by the students' English skills when they enter a program.

Educators from the CELLA consortium states, as well as AW, insisted on a valid measure of English proficiency for ELL students at all skill levels. They wished for the instrument to provide highly accurate results for students who did not fit conventional notions about which English skills ELL students might possess at a given grade level. Since the population to be tested may not reflect the population in a particular state, district, school, or individual

[1] They may speak a first language other than English or they may speak a nonstandard dialect of English. It is also possible that they were born in the United States, but speak a variety of *immigrant English* or l*earner English* that they have learned from their non-native English-speaking friends (Scarcella, 1996).

RETURN to TABLE of CONTENTS

classroom. The consortium educators viewed accurate results as particularly important in the areas of reading and writing, skills that tend to develop more slowly than oral language skills for ELL students. An assessment of high-level English skills at the high school level may be quite necessary for determining exit from ESL or bilingual programs, but it can also be enormously frustrating for students at the initial stages of English development, and at the same time provide little formative assessment information of value. Finally, the consortium states wished the assessment to provide a rigorous instrument for measuring the *growth* that students achieve in their English skills, since they believed that the extent of such growth over the course of an academic year is a more useful indicator of the contribution of an ESL or bilingual program than how many students attained proficiency in any particular year.

The design of CELLA addresses these legitimate interests of educators who are involved in serving ELL students every day. CELLA employs a functional approach to assessing reading and writing that accurately measures the wide range of skills of ELLs described above without extending testing time or frustrating students with inappropriate items. In the reading and writing domains, there are four test levels of the CELLA, Levels A through D that reflect the range of English skills that ELLs must master to succeed in grades K–12. A student at the high school level (grades 9–12) may take any one of the four levels, depending on how far along he or she is on the continuum of mastering the necessary skills. Similarly, a student in the middle grades (6–8) may take any one of three levels, A through C, while a student in the upper elementary grades (3–5) may take either of two levels, A or B. All students in the early primary grades (K–2) take the same level, A. The item prompts in levels A, B and C were developed to be as age neutral as possible, so they are appropriate for students at different ages. Because the tasks administered to each student are appropriate to his or her instructional level, the detailed information provided is diagnostically relevant for every student. (Additional information on diagnostic reports is provided later in this chapter.)

Use of Research in Identifying Objectives and Designing Tasks

In identifying the objectives to be assessed, the types of tasks to used, and the type of information that would result from the assessment, the development of CELLA took into account the research on ELLs reported in the literature. One important general principle was a strong emphasis on academic English.

Academic English involves mastery of a writing system as well as proficiency in reading, speaking, and listening. Experts agree that it entails the body of knowledge, strategies and skills necessary for students to participate in school activities and learn from content instruction. An important component of academic English is the system of sounds used in the language. To use academic English, learners must learn the phonological features of academic English, including stress, intonation, and sound patterns. For example, when learners are exposed to new academic words such as *manipulate* and *manipulation*, they must learn their distinct stress patterns.

Another important component of academic English is vocabulary. To communicate in school situations, English learners must not only develop everyday words, they must also develop more challenging, academic words, as well as content-specific words (Biemiller, 2001; Beck, McKwon & Kukan, 2002). They must learn that words used in everyday conversational English take on special meanings in academic English. Consider the words *fault, power, force, active,* and *plate*. Even common words can take on very precise and possibly unfamiliar meanings in classroom settings. For example, the use of the preposition *by* to mean *according to* in the sentence, “I want them to sort by color” (Bailey, Butler, LaFramenta & Ong 2004).

The grammatical component of academic English entails sentence structure and morphology (the structure and form of words in a language, including inflection, derivation, and the formation of compounds); complex sentences, such as passive structures (“The book was written by Shakespeare”); and conditionals (“If someone showed you a set of 45 dots next to a set of 5 dots, then you’d probably be able to tell right away which set has more dots”). Those who have mastered academic English also know how to use the entire gamut of modal auxiliaries (e.g., *will/would, can/could, may/might, should, must, have to,* and *ought to*) and not just those which occur frequently in everyday English, such as can and would.

Other critical features of academic English include those that enable English learners to signal levels of politeness and formality, to establish their credibility in school contexts, and to communicate coherently. In reading, knowledge of such features helps students to gain perspective on what they read, to understand relationships, and to follow logical lines of thought. In writing, such features

RETURN to TABLE of CONTENTS

help students develop topic sentences and provide smooth transitions between ideas.

Meta-linguistic skills are also essential to the development of academic English. These skills make it possible for students to take language apart, enabling them to evaluate reading passages critically and edit writing effectively. This is why editing skills are assessed on the CELLA.

Academic English is also more abstract and de-contextualized than informal English[2] —it has fewer contextual supports to help students understand and communicate. In informal English, students need not rely only on language in order to comprehend and reply to their interlocutors; instead, they can use contextual supports such as body language and intonation to express themselves. They can observe others' nonverbal reactions (e.g., facial expressions, gestures, and eye movements); interpret vocal cues, such as phrasing, intonation, and stress; observe pictures, concrete items, and other contextual cues; and ask for statements to be repeated and/or clarified. In academic English, they cannot rely on nonverbal contextual cues to figure out what others are saying. Non-verbal clues are absent; there is less face-to-face interaction; the language used is often abstract; and they lack the English proficiency and background knowledge to understand its accurate use.[3]

In CELLA, academic English is assessed in all four domains—reading, writing, speaking and listening. Hence, CELLA can be used to exit ELLs into programs in which a high level of English proficiency is assumed.

Research was also applied in the foundational components of reading—phonemic awareness, decoding, oral fluency, vocabulary, and comprehension. These components, outlined in the National Reading Panel Report (1999), are essential in developing reading in English as a second language as well as a first language (reported in August & Shanahan, 2006). In accordance with research, the reading assessment items were designed to measure phonological processing, letter knowledge, and word reading, all valid measures for assessing the reading skills of ELLs who are just beginning to learn to read. These measures can be used to identify ELLs who are likely to benefit from additional instruction.[4] These measures are categorized as follows:

1. Measures of phonological awareness, such as segmenting the phonemes in a word, sound blending, and rhyming
2. Measures of familiarity with the alphabet, especially measures of speed and accuracy in letter naming
3. Measures of reading single words and knowledge of basic phonics rules
4. Measures of reading connected text accurately.

Further, because vocabulary is essential to reading development,[5] CELLA assesses vocabulary at all levels. In developing reading items, the critical importance of reading comprehension and higher order thinking was also taken into account. Many of CELLA's reading assessment items are quite challenging; they require complex, analytical thought and the effective orchestration of comprehension strategies.

The approach toward reading assessment adopted by CELLA was affirmed in a recent landmark report summarizing the research on instruction and assessment of English language learners. The U.S. Department of Education's Institute of Education Sciences (IES) created the National Literacy Panel on Language-Minority Children and Youth to identify, assess, and synthesize the best research on the education of language-minority students with regard to literacy attainment. The panel established strict criteria for selecting research of high quality and rigor. The panel's findings are consistent with the main features of CELLA's design and affirm the research base that was used in its development.

The National Literacy Panel concluded that instruction providing substantial coverage in the essential components of reading—identified by the panel (NICHD, 2000) as

[2] Schleppegrell, 2004; Butler, Bailey, Stevens, Huang & Lord, 2004; Scarcella, 2003; Echevarria & Short, 2002; Cummins, 1979

[3] August, 2003; Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2006). Educating English language learners: A synthesis of research evidence. New York: Cambridge University Press.

[4] Arab-Modhaddam & Senechal, 2001; Chiappe, Siegel & Gottardo, 2002; Chiappe, Siegel, & Wade-Woolley, 2002; Geva, Yaghoub-Zadeh, & Schuster, 2000; Lesaux & Siegel, 2003; Limbos, 2006, Limbos & Geva, 2001; Manis, Lindsey, & Bailey, 2004; Quiroga, Lemos-Britton, Mostafapour, Abbott, & Berninger, 2002; Verhoeven, 1990, 2000.

[5] August, 2003; Carlo et al., 2004; Biemiller, 2001; Beck, McKwon & Kukan, 2002; Nagy, 1988.

RETURN to TABLE of CONTENTS

phonemic awareness, phonics, fluency, vocabulary, and text comprehension—has clear benefits for language-minority students. The panel also found that current assessment tools should include measures that can help predict a student's literacy acquisition over time. CELLA provides educators with specific information regarding student performance in all of the essential aspects of English acquisition identified by the panel. The panel also reviewed the area of student assessment and found that most of the traditional assessments do a poor job of gauging particular strengths and weaknesses, but that such diagnostic measures are essential when evaluating an English language learner.

Needs of Local Users

Another major objective of CELLA was to address the assessment and instructional needs of all ELL students in a wide range of classroom settings. ELLs can be found in a variety of instructional programs; therefore, assessments must be appropriate to the range of programs, including bilingual, English-only, and English Language Development/English as a Second Language programs. For this reason, CELLA was designed to be relatively easy to administer and score, not only by ESL experts, but also by individuals lacking knowledge of ELLs' needs and experiences, as well as their educational, cultural, and linguistic backgrounds.

The educators in the consortium states consistently pushed the test designers to shorten the time required for test administration to the minimum necessary. There were some constraints, however. The federal requirement that the assessment measure students' speaking skills (in addition to the more manageable listening, reading, and writing skills) necessitates that part of the test be individually administered. Similarly, valid measurement of initial language acquisition (Level A) requires a small number of individually administered listening and reading items, along with other group-administered listening and reading items. Despite these administrational challenges, the final product is a highly efficient instrument. Estimated test administration times for Level A are one hour for students in grades K–1 (of which only 15 minutes is individually administered at grade 1) and one and a half hours for students in grades 2–12 (of which, again, only 15 minutes is individually administered). Test administration time for all parts of Levels B, C and D is slightly more than two and a half hours (with individually administered sections of 12 to 14 minutes).

Even though CELLA is efficient, the test is still able to meet another goal: providing useful diagnostic information to local users. CELLA accomplishes this with powerful diagnostic reports gleaned from student results. While federal law requires that CELLA assess comprehension (a measure that is provided on CELLA subscore reports), every level of CELLA also addresses vocabulary acquisition by providing information on students' vocabulary skills, which research has shown is of great importance. Further, levels B through D of CELLA distinguish between oral vocabulary and written vocabulary, which may be quite different for individual students and suggest different types of instructional intervention.

At levels B through D, CELLA result reports distinguish between student skills in grammar, editing, constructing sentences, and constructing paragraphs. At the same levels, CELLA Speaking section reports disentangle student skills in vocabulary, accurate and appropriate question asking , and extended speech (e.g., expressing an opinion or discussing information in a graph).

CELLA designers drew on the wealth of scientific based research on reading instruction to develop the score reports for students at the stage of initial language acquisition (Level A). CELLA Level A reports distinguish between decoding, vocabulary, and comprehension results and, for students at grades K–1, basic print concepts (e.g., direction of print, names of letters). Further, the test administrator is directed to record useful information on students' oral reading fluency, including rate and accuracy.

These and other results, available for individual students from a CELLA administration, provide valuable diagnostic data from carefully constructed and validated tasks based on the best research and the judgment of experienced educators.

The Use of Standards as a Basis for Test Development

In order to develop an assessment consistent with the highest professional standards at every stage of design and development, ETS employed its model of evidence-centered design (ECD) for test development. ECD is an intellectual discipline for the design of assessments and is defined by three components: claims, evidence, and tasks. Prior to designing an assessment, the test development team asks three questions:

RETURN to TABLE of CONTENTS

1. What claims will be made about student performance on this test?
2. What evidence is needed to support those claims?
3. What tasks can be designed to gather the evidence needed to support those claims?

In the case of CELLA, the claim to be made is that a student has the requisite English-language skills to succeed academically in the mainstream U.S. classroom or is making progress toward attaining those skills. The evidence for those claims is found in the student achieving the objectives that are defined in the CELLA Proficiency Benchmarks, a matrix of component skills at each grade level that students are expected to attain. The tasks that gather the evidence are the items of various types found in CELLA.

The first stage in the development planning process was the creation of a common set of assessment objectives acceptable to the five states in the CELLA consortium. The CELLA Proficiency Benchmarks were developed based on the experience and professional judgment of AW researchers, language researchers, and ETS test developers. The benchmarks were reviewed and approved by educators and other representatives of the five states. The CELLA Proficiency Benchmarks act as a set of common assessment objectives for ELLs from the various states.

A important step in the development of the CELLA benchmarks was a careful alignment analysis of the benchmarks with the standards of consortium states. Three of the five consortium states had developed ELP standards by this stage of the process and provided them for this purpose. AW—with substantial organizational experience in this type of analysis—performed this work with the assistance of similarly experienced content experts. As a result of this process, high alignment was documented between the CELLA benchmarks and state ELP standards for the participating consortium members, as these existed at the time of the analysis. (While the focus was on alignment to state ELP standards, a similar alignment analysis was also performed for informational purposes with respect to consortium states' mainstream academic reading/language arts standards.)

In addition to the overall judgments which found high levels of alignment, the analysis also provided quantifiable results regarding alignment to more detailed state objectives under each state's education standards. Alignment at this more detailed level was also quite high. The CELLA benchmarks demonstrated alignment to 100% of the Florida ESOL standards and to 90% of the detailed state objectives. Similarly, the CELLA benchmarks demonstrated alignment to 100% of Michigan's ESOL standards and to 95% of the detailed objectives. For Pennsylvania, the CELLA benchmarks demonstrated alignment to 100% of the state ESOL standards and to 89% of the detailed objectives.

Comparison Between New and Prior Generations of ELP Tests

Previously, ELP tests were used primarily for placement into English language instruction and generally focused on the lower and intermediate proficiency levels. Among the new generation of ELP assessments, the Comprehensive English Language Learning Assessment (CELLA), differs from previous ELP tests in that it:

- focuses on assessing the specific language skills needed to learn effectively in grade-level content classes taught in English
- is standards-based, with the CELLA proficiency benchmarks acting as a shared set of assessment objectives, developed in alignment with state standards for English learning proficiency
- was designed as a comprehensive, summative assessment of ELLs' skills at a given point in their academic career; it is far more than a test used for quick placement of newly arrived ELLs, though the information derived from it may also be used for this purpose
- includes direct measures of the productive skills of speaking and writing
- is fully NCLB compliant, with separate measures of listening, speaking, reading, and writing as well as reporting a wide range of subscores, including comprehension and vocabulary

Creating the Test Blueprint

The next stage in the process was the creation of *test blueprints* for each of the four CELLA levels. The blueprints are the design specifications that describe the different categories of content that appear in the test forms and how that content is distributed. For example, the CELLA Listening sections contain items that test vocabulary and comprehension. At Levels B through D the vocabulary items make up 35% of these items, while the comprehension items make up the remaining 65%. At Level A, the breakdown is 25%

RETURN to TABLE of CONTENTS

and 75%, respectively, for vocabulary and comprehension items in the listening section.

The test blueprints for CELLA were developed hand-in-hand with the proficiency benchmarks. Drafts were created by AccountabilityWorks and ETS and refined through a process of review and comment by representatives of the consortium states.

The final stage of the development planning process was the creation of *item specifications,* which are descriptions of the tasks students will perform to provide evidence of their proficiency. Item specifications show in detail *how* the evidence about student proficiencies is captured by the items on the test. In addition, they provide models and examples to test developers so that future test forms will be created to collect comparable information.

One of the greatest challenges in designing a four-domain ELP assessment is to create a test that is valid and reliable but not overly time-consuming for students and teachers. While lengthening a test will result in more reliable test scores, this benefit must be balanced against the need to keep testing time to a minimum to avoid overburdening students and teachers. The CELLA blueprints and item specifications were finalized with this challenge in mind—to create the most reliable test possible without overtaxing students and frustrating teachers. The test blueprints and item specifications were reviewed and approved at a meeting of representatives of the consortium states in the spring of 2004.

Item Development

Once the test design had been completed—meaning that the proficiency benchmarks, test blueprints, and item specifications had been created and approved—development of the CELLA test items began.

The items used in CELLA were initially created by experienced ESL professionals and reviewed internally by ETS test development staff to assure that each item would assess the intended standard, have strong technical quality, and be free of any bias or sensitivity issues that might affect student performance. After the ETS reviews, all CELLA items were reviewed and approved for content and fairness by AW researchers and language experts as well as by committees of educators from the consortium states.

The consortium participants recognized that the life experiences of ELLs differ widely—not only from those of native speakers but also from the experiences of ELLs of other cultures. Thus, ELLs cannot be assumed to share common experiences outside of school. For this reason, CELLA item content—whether in items assessing academic language or social/interpersonal language—is primarily in the context of the U.S. K–12 school system. A student might be asked to listen to a short announcement about an upcoming field trip; speak about a graph showing what hobbies are most popular at a particular school; read academic passages that might appear in a U.S. social studies, science or language arts textbook; or write a letter to the editor of the school newspaper on a school-related topic.

Item Types

In designing CELLA items, ETS drew on its wide range of experience as well as new innovations to create item types that are valid and reliable for the assessment of K–12 English language learners. This included item types that are suited to assessing not only students' general English language proficiency, but also those specific language skills that enable K–12 ELLs to learn grade-level content and succeed in English-medium classrooms.

For example, the CELLA Listening section includes item types that are specifically designed to measure students' readiness for the mainstream classroom. The "Teacher Talk" items require students to listen to a teacher providing instructions or making an announcement in a classroom, measuring a student's ability to comprehend school-based information. The "Extended Listening Comprehension" items present oral academic material followed by three questions. They are carefully designed so that any content-area vocabulary they contain is clearly explained within the stimulus. In this way, the items assess students' ability to learn new concepts and information in a content area while remaining fully accessible to linguistically proficient students who have not yet had instruction in the item's content.

The CELLA Speaking section consists entirely of constructed-response items ranging from directed-response "Oral Vocabulary" items (scored as right or wrong), to controlled-response "Speech Functions" and "Personal Opinion" items (scored on a 0–2 rubric), to more extended-response "Story Retelling" and "Graph Interpretation" items (scored on a 0–4 rubric). Of these, some of the more innovative item types are Speech Functions, which assesses a student's ability to construct questions that are functionally appropriate and grammatically accurate, and Graph Interpretation, which requires students to analyze a

RETURN to TABLE of CONTENTS

graph or chart and describe the information presented. The Graph Interpretation items are carefully designed to assess students' abilities to talk about numerical information and make comparisons, while not requiring that students have had grade-level instruction in mathematics.

The item types for the CELLA Reading and Writing sections have been designed to assess literacy skills across the entire spectrum, from an understanding of basic print concepts to full reading comprehension and writing skills at grade level. Functional-level testing of reading and writing assures that students, regardless of their age or grade level, are administered only item types designed to effectively measure their current level of literacy in English.

At Level A, items at the lowest skill level assess students' understanding of print concepts, phonemic awareness, and ability to recognize and name letters. Additional items assess recognition of sight words and decoding skills. At the higher end of Level A, students are asked to read independently and comprehend brief passages. Additional, more challenging passages and items are administered to students in grade 2 and above. At Levels B through D, the CELLA Reading section contains a handful of discrete items that test knowledge of grade-level vocabulary. The bulk of each reading section at these levels consists of grade-level appropriate academic or literature-based passages. Each passage is followed by four to six items that test such reading skills as recognizing main ideas, identifying details, making inferences and predictions based on the text, understanding vocabulary in context, and comprehending cohesive elements. Students taking CELLA at Level A or B also complete a timed "Reading Aloud for Fluency" task that assesses their reading rate, accuracy, and ability to read words meaningfully.

A combination of multiple-choice (MC) and constructed-response (CR) item types are used in the CELLA Writing section. At Level A, all of the items are in the constructed-response format, testing beginning writers' ability to write dictated letters, words, and sentences, and construct original sentences based on pictures. Constructed-response items in Levels B through D require students to write individual descriptive sentences and questions as well as paragraphs. In each test form, students write two paragraphs requiring the use of different grade-appropriate rhetorical structures such as narration, description, persuasion, and compare-and-contrast discussion. Multiple-choice item types are used to assess students' understanding of grammar and structure, including the ability to recognize errors and use transitional devices. All CELLA multiple-choice writing items are presented in the context of a paragraph appropriate to reading level.

Assessing Comprehension

Title III instruments are required to assess children's proficiency in comprehension as well as in listening, speaking, reading, and writing. In the development of CELLA, care was taken to ensure that comprehension is assessed both in listening and in reading and that comprehension scores are reported in the way that provides the most detailed and useful information possible for students, teachers, and parents.

Comprehension on CELLA is defined as those items within the listening and reading sections that require comprehension of extended texts. In the listening section this includes "Short Talk" and "Extended Listening Comprehension" items (and excludes items where the stimulus is not longer than a single sentence). In the reading section, comprehension includes "Reading Comprehension" items (and excludes items that focus exclusively on vocabulary).

In each form and test level of CELLA, comprehension is reported not as a single score, but as a reading comprehension subscore and one or more listening comprehension subscores. (At Level A there is a single listening comprehension subscore; at Levels B through D there are two listening comprehension subscores: for "Short Talks" and for "Extended Speech.") These more detailed comprehension subscores reflect a stated purpose of CELLA to provide as much diagnostic information as is possible within the confines of a large-scale standardized assessment.

CELLA comprehension subscores are included on individual student reports and are presented as raw scores (i.e., points achieved out of points possible). Because of the detailed nature of the CELLA comprehension subscores, there are currently no aggregated reports that include comprehension scores.

Functional Level Testing

As noted above, English language learners often demonstrate a wide range of English language proficiency levels that do not necessarily correlate with their grade level. This disparity between grade level and functional level presents a major challenge when creating assessments for ELLs. Such assessments must not only measure the relatively high proficiency levels needed to inform reclassification decisions (i.e., proficiencies high enough for students to

RETURN to TABLE of CONTENTS

succeed in English-speaking classes in the content areas), they must also provide sufficient information about student performance at the lower proficiency levels.

CELLA addresses this challenge through the use of functional level testing of reading and writing.[6] The reading and writing sections are organized into four functional levels based on language proficiency:

Level A: Initial literacy skills

Level B: The application of literacy skills toward the development of new knowledge

Level C: More advanced applications of literacy skills toward the development of new knowledge

Level D: Literacy skills necessary for success in higher education or the workforce

In cases where prior CELLA assessment data is not available for students, a short, easy-to-score *locator* test is available to determine the appropriate test level of reading and writing sections to administer to individual students.

Each of the four functional levels addresses the *highest level* of language proficiency necessary to succeed in the English-speaking classroom at a specific grade span. *Together with any lower levels*, they form the full range of the CELLA assessment appropriate for ELL students at any given grade.[7] Thus, ELL students taking the CELLA Reading and Writing sections may be administered the following test levels according to their grades:

Grades 9–12:	Levels A through D
Grades 6–8:	Levels A through C
Grades 3–5:	Levels A through B
Grades K–2:	Level A

Vertical scales, which place all of the items in each domain on the same scale, allow comparison of performance on one level of the test to performance on another.

The use of functional level testing allows CELLA to provide accurate information about all students, regardless of the level of development of their literacy skills. Additionally, administration of the appropriate functional level test will decrease frustration for students who might otherwise be required to take an assessment that is not appropriate for them.

Field Testing

Students in kindergarten through grade 12 from the five consortium states participated in the field test. Each state was asked to select students from the full range of proficiency levels. There were three field-test forms at each level, from which the most appropriate items were selected to construct two operational forms. Each field-test form contained two types of items: items designed for the specific test level and items that enabled vertical linking. In each test level, these three field-test forms contained common items that were also used to link the forms horizontally.

Table 1. Number of Students in the Field Test by State and Grade Range

	Grade Range			
State	K, 1, 2, 3	4, 5, 6	7, 8, 9	10, 11, 12
Michigan	859	691	711	640
Maryland	1,273	835	942	750
Florida	1,348	1,003	1,035	945
Pennsylvania	1,233	749	709	585
Tennessee	711	444	369	349
Total	5,424	3,722	3,766	3,269

Item analyses (IA) were conducted to establish the reasonableness of keys and the reliability of each field test. To assess the difficulty of each multiple-choice item, the proportion of students correctly answering the item (called the *p*-value) was used as the index of item difficulty; for constructed-response items an analogous index consisting of the mean item score divided by the maximum possible item score was used. The correlation between students' item scores and their total test scores was used to assess item discrimination for both item types. These field-test item performance indices were used in operational form construction. Table 2 contains the means and standard deviations of the item difficulties and item discriminations for the final test forms.

To evaluate internal consistency of the field-test forms, Cronbach's alpha was used for the listening and reading test sections while a stratified coefficient alpha was used for the writing and speaking test sections.

[6] Because ELLs generally develop oral skills more rapidly than literacy skills—and because of logistical challenges to administering different levels of listening to a single class—the speaking and listening sections are administered to students according to their grade level.

[7] To the extent possible, items in levels A through C of the Reading and Writing sections were designed to be age-neutral, thus appropriate for students at a range of ages.

RETURN to TABLE of CONTENTS

Forms Construction

Operational Test Forms

Operational forms were created by attempting to keep the highest quality field-test forms intact. Items were swapped in and out of the intact form only as necessary to balance the difficulty and content representation of item types. Items for the second operational form at each level were selected so that the difficulty and discrimination of the items matched those of the first operational form. An effort also was made to keep the operational item order very close to what it was in the field-test form. Items were sequenced within form from easy to difficult, when possible. Both operational forms were assembled according to the test blueprint. Table 2 contains the means and standard deviations of the item difficulties and item discriminations for the final test forms.

Vertical Scaling

The items calibrated by test level were vertically scaled using the test characteristic curve (TCC) method described by Stocking and Lord (1983). Because vertical scaling produces a continuous scale of item difficulty expressed in terms of test takers' ability (i.e., the theta metric), it enhances score interpretability across the four test levels. For vertical scaling of CELLA, the Level B items were used to define the base scale for vertical scaling. That is, the items at the other three levels were placed on an ability scale in relation with the Level B items. The ability metric then was transformed to a 3-digit scale score metric to make scale scores more familiar and comprehensible to test users. The range of scale score for Listening/Speaking from Level A to Level D is 495 to 835. The Reading scale ranges from 345 to 820, and the Writing scale ranges from 515 to 850.

Scale Anchoring and Standard Setting

In order to increase the interpretability of the CELLA scale scores, descriptions of students' knowledge and skills at different points on the scale were developed. This was achieved through a process of scale anchoring. This included identifying selected scale score anchor values throughout the range of performance. Students who performed near these anchor values on the field test were identified. Based on the performance of these students, specific items were identified on which these students usually were successful. Content experts reviewed these items and developed behavioral descriptions that highlighted the major behaviors typically evidenced by students at each anchor value. These descriptions can help students, parents, and teachers understand the meaning of student scale scores. Further, they can be useful to states developing performance standards for the CELLA (i.e., standard-setting) for various state purposes. The CELLA scale anchoring was very successful; it showed a progression of what students can do at each of the anchor values for the three vertical scales. This progression can also be viewed as one piece of evidence of the validity of the CELLA vertical scales.

The scale anchoring process differs from a formal standard-setting process and does not replace it. In standard setting, expert committees are given performance-level descriptions (e.g., for *proficient* or *advanced* performance) and information about test items, then tasked with identifying the test performance that is required, according to their judgment, to reach each performance level. The scale-anchoring procedure adds behavioral meaning to the scale scores provided by CELLA; it is *not* intended to replace states' individual formal standard settings, which define student expectations in each state. The consortium determined that it would be inappropriate to force consensus on such policy matters across the participating states. The result is that states outside the CELLA consortium may find it a more flexible instrument for their purposes.

The Tennessee Department of Education administered the test and set standards for CELLA in 2006. They identified three proficiency levels of CELLA for each grade level and skill area. Because there are far fewer students taking CELLA in higher grade levels than lower grade levels, the cut scores for grades 9 through 12 within each skill area were the same, while the cut scores for grades K through 8 were different for each grade level. The Tennessee standard setting provides a good illustration of how different circumstances and policy preferences of individual states may affect how the standards can be set.

For more information about the development of the final test forms and the locator test, the CELLA Technical Summary Report summarizes the field testing, calibration, and vertical scaling of test items.

Test Validation

As described earlier, CELLA was developed following a rigorous test development procedure in order to ensure the validity of the test. State ESL standards and teachers were heavily consulted to align the test content to local expectations of English language acquisition. During the

RETURN to TABLE of CONTENTS

Table 2. Final Form Summary Statistics

			Difficulty		Discrimination	
Domain	Test Level/ Form	No. of Items	Mean	SD	Mean	SD
Listening	A1	20	0.74	0.13	0.62	0.09
	A2	20	0.74	0.13	0.57	0.09
	B1	22	0.72	0.13	0.54	0.09
	B2	22	0.69	0.15	0.54	0.07
	C1	22	0.73	0.12	0.61	0.09
	C2	22	0.71	0.13	0.60	0.08
	D1	22	0.76	0.14	0.58	0.09
	D2	22	0.75	0.15	0.55	0.09
Speaking	A1	10	0.72	0.14	0.75	0.08
	A2	10	0.68	0.14	0.74	0.08
	B1	14	0.72	0.13	0.75	0.07
	B2	14	0.68	0.13	0.74	0.08
	C1	13	0.69	0.11	0.74	0.13
	C2	13	0.67	0.11	0.75	0.12
	D1	13	0.72	0.11	0.72	0.13
	D2	13	0.74	0.07	0.71	0.13
Reading	A1	21	0.70	0.16	0.64	0.13
	A2	21	0.70	0.16	0.63	0.16
	A1-Ext	31	0.64	0.17	0.58	0.14
	A2-Ext	31	0.66	0.16	0.61	0.14
	B1	27	0.60	0.16	0.59	0.14
	B2	27	0.61	0.13	0.60	0.10
	C1	26	0.57	0.14	0.59	0.11
	C2	26	0.57	0.13	0.58	0.09
	D1	26	0.62	0.13	0.57	0.10
	D2	26	0.63	0.12	0.57	0.08
Writing	A1	7	0.50	0.13	0.84	0.05
	A2	7	0.47	0.12	0.80	0.01
	A1-Ext	16	0.58	0.14	0.77	0.12
	A2-Ext	16	0.57	0.14	0.76	0.10
	B1	25	0.60	0.11	0.58	0.14
	B2	25	0.61	0.11	0.59	0.15
	C1	25	0.62	0.13	0.57	0.19
	C2	25	0.59	0.14	0.58	0.17
	D1	25	0.63	0.14	0.65	0.12
	D2	25	0.67	0.11	0.61	0.12

RETURN to TABLE of CONTENTS

test development, all items were evaluated in terms of bias in favor of or against a particular group of test takers. Such bias represents irrelevant factors on the test. Differential (DIF) analyses compared item performance of male students versus female students. In total, five items were flagged for DIF. These items were reviewed by content experts to determine if the items contained inappropriate content. Three of the items were dropped from the pool, and two of the items were retained in the item pool and used in the final test forms that were subsequently developed. At the test level, CELLA displays no gender-related bias, the result of which contributes to the validity argument.

Additional evidence for the validity of CELLA as a test of a language proficiency is found in the factor analysis (Fitzpatrick et al., 2006) . The four domains of CELLA were analyzed for a factor structure. High correlations among four domains (reading, writing, listening, and speaking) and one factor were observed. The results are desirable given that the test is supposed to measure the single construct of language proficiency in four domains. Interestingly, the two factor structure that separated reading and writing from listening and speaking were also found. In general, there was no evidence indicating that the tests tap other dimensions than language ability.

Since CELLA is relatively new, it will take some time to find validity evidence in the field. One area to look into will be whether the test does what it claims, such as whether CELLA can identify those who need additional language support until they are fully proficient to function in mainstream classrooms. Further research may investigate how effectively the test differentiates ELLs who need additional support from those who don't. Analysis of some longitudinal data, with the help of local education agencies, will also be helpful to determine the use of CELLA as a test that predicts the linguistic readiness of ELLs for learning academic content, as well as a test that is sensitive enough to measure linguistic growth.

Test Accommodations

CELLA is not intended to be a timed test; thus test administrators are encouraged to give students sufficient time to complete the test without feeling rushed. Students with Individualized Education Plans (IEPs) can be accommodated in whatever way their IEP calls for, as long as the accommodation does not conflict with the test construct. For example, a student would not be allowed to use an English-Spanish dictionary because CELLA is a test of English proficiency.

For students who are visually impaired, CELLA is available in Braille and large-print versions. (On the braille version of test, items that include pictures and cannot be brailed are not included in the students' score.) Students who are deaf or severely hard of hearing and do not use oral language may be exempted from taking the CELLA Listening and Speaking sections.

Test Administration

In broad terms, CELLA is divided into four sections (listening, speaking, reading, and writing), each of which is administered in a separate session. Individual schools can schedule these sessions at their discretion. Because the design of Level A (grades K–2) differs substantially from the design of Levels B through D (grades 3–12), these two grade spans are presented separately below.

Level A

It is recommended that all sections of the Level A test be administered individually to students in kindergarten. For grades 1 and 2, listening, reading and writing sections can be administered in small groups. (Teachers should to use their discretion to create groups of a manageable size.) The one-on-one section includes not only speaking items but also those listening and reading items that require individual administration.

The development of initial English literacy skills is very different for students at the beginning of kindergarten compared to students in grades 2 and higher. For that reason, Level A reading and writing sections are divided into two parts. The core of Level A (without the extension) assesses skills at the most basic level and is appropriate for administration in grades K and 1. Level A Extension contains more challenging reading comprehension and writing items and is appropriate for students at the initial stage of English literacy development in grades 2 through 12. Students in kindergarten and grade 1 take only the Level A core items. Students in grade 2 through grade 12 take the Level A core in its entirety and then continue on to take the Level A Extension.

The times required for administration of the Level A and Level A Extension tests are shown in Table 3.

RETURN to TABLE of CONTENTS

Levels B through D

For Levels B through D (which may be administered in grades 3–12), the listening, reading, and writing sections are all group-administered assessments, with a recommended total testing time of approximately two and a half hours. The speaking section is administered individually to each student. The duration of the speaking section is estimated at 12 to 14 minutes, depending on the proficiency level of the student being tested. Table 4 shows the times required for administration for the test sections at Levels B through D.

Scoring

The Level A listening and reading sections contain only multiple-choice items, and students mark their answers directly on a scannable test book which is returned to ETS for scoring. At Levels B through D, the listening and reading sections also contain only multiple-choice items, but students bubble in their responses on answer documents which are returned to ETS for scoring.

At all levels, the CELLA Writing sections require students to compose original texts. At Level A, students progress from writing down dictated words to composing

Table 3. Level A

	Level A		Level A + Extension		Administration*
	Items	Time	Items	Time	
Listening (MC)	15	15	15	15	individual/group
Reading (MC)	15	15	25	35	individual/group
Writing (CR)	7	15	16	30	individual/group
One-on-One (CR)**	21	15	21	15	individual
Total Items	58		77		
Total Time (estimated)	1 hr.		1 hr. 35 min.		

** Level A and its Extension have been designed for administration to small groups (i.e., 6 to 8 students) wherever possible for grades 1 and up. Individual administration of all test sections is recommended for students in kindergarten.*

*** In addition to speaking items, the one-on-one section includes those listening items and reading items which must be individually administered.*

Table 4. Levels B-D

	Level B		Level C		Level D		Administration
	Items	Time	Items	Time	Items	Time	
Listening (MC)	22	30	22	30	22	30	group
Speaking (CR)*	14	14	13	12	13	12	individual
Reading (MC)	26	45	26	45	26	45	group
Writing (MC and CR)**	25	70	25	70	25	70	group
Total Items	87		86		86		
Testing time***	2 hrs. 39 min.		2 hrs. 37 min.		2 hrs. 37 min.		

** The speaking section is slightly longer at Level B because it includes administration of the Reading Aloud for Fluency item type.*

***The writing section contains 19 multiple-choice items and 6 constructed-response items (4 sentence writing tasks and 2 paragraph-writing tasks).*

**** An additional fifteen minutes are estimated to be necessary for student entry of personal/demographic information.*

RETURN to TABLE of CONTENTS

original sentences, and they write directly in the test books. At Levels B through D, students write original sentences and paragraphs based on written prompts. At Levels B through D, the student writing samples are captured on the student's answer document.

Scoring of Constructed-Response Writing Items

The CELLA Writing sections contain both multiple-choice items and constructed-response items requiring students to provide a written response to a prompt. (Level A contains only constructed-response items.) For all test levels, students' written responses can be scored centrally by ETS. The handwritten responses are scanned to create secure electronic files that are read by raters over the Internet. ETS employs qualified educators who are trained to rate specific items. Using the appropriate rubrics for each item, raters are calibrated against anchor responses for each score point. (The anchor and training responses were gathered during the CELLA field test.) When raters are sufficiently calibrated, they begin scoring live essays.

For states that prefer to score written responses locally (either when CELLA is used as an annual summative assessment or when it is used for placement purposes), complete *Scoring Guides for Writing* have been created for each test level and form. These guides include explanations of the CELLA constructed-response item types and the scoring process, as well as rubrics and authentic student responses that have been selected as anchor and training responses for awarding each score point.

After students' written responses have been assigned scores, those scores are electronically merged with the students' listening, reading, and speaking scores that were scanned from their respective sections of the answer document.

Scoring of Speaking Responses

The scoring of speaking responses occurs in the moment—that is, the test administrator who sits down with the student and asks the questions is the same person who rates the student's responses. The scores are marked directly on the student's answer document by the test administrator.

In order to ensure accurate and consistent scoring, all administrators must undergo scoring training before administering the CELLA Speaking section. Training for administering the CELLA Speaking section is based on the *Scoring Guide for Speaking* that has been prepared for each of the four test levels. Each scoring guide contains items comparable to those in the speaking section, rubrics for each item type, an audio CD of authentic student responses that have been selected as *anchors* and *trainers* for awarding each score point, and transcriptions and rationales for each recorded response. An answer key for the sample training items is also included.

Training can be conducted individually (as self-study) or in a moderated group. Individual training takes approximately two and half hours. Group training can be expected to take somewhat longer as time must be allowed for discussion. The training CD includes a broad range of examples of actual student responses so that teachers (or other test administrators) have an opportunity to practice making score determinations. Educators who have reviewed the training materials have found the student responses to be highly realistic, with incorrect responses very representative of typical student errors.

CELLA Today

As of this writing (late fall 2006), CELLA has been administered on three separate occasions to ELL students, resulting in nearly 300,000 individual test administrations. CELLA was administered in spring 2005, spring 2006, and fall 2006. CELLA products and services are available to states. These include two operational forms for each of four levels (A through D), a short locator test for determining the correct level to administer to students lacking prior CELLA data, administration manuals, training materials, and scoring services. Those interested in learning more about CELLA may contact AccountabilityWorks.

References

August, D., & Shanahan, T. (Eds.) (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on language minority children and youth*. Mahwah, NJ: Erlbaum.

Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C. (2004). *Towards the characterization of academic language in upper elementary science classrooms* (CSE Tech. Rep. No. 621). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

RETURN to TABLE of CONTENTS

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust vocabulary instruction*. New York, NY: Guillford Press.

Biemiller, A. (2001). *Teaching vocabulary: Early, direct, and sequential*. American Educator. http://www.aft.org/pubsreports/american_educator/spring2001/vocab.html

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D., Lively, T., & White, C. (2004). Closing the gap: Addressing the vocabulary needs for English language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, *39*, 188–215.

Chiappe, P., & Siegel, L. S. (1999). Phonological awareness and reading acquisition in English- and Punjabi-speaking Canadian children. *Journal of Educational Psychology*, *91*, 20–28.

Chiappe, P., & Siegel, L. S. (2006). A longitudinal study of reading development of Canadian children from diverse linguistic backgrounds. *Elementary School Journal*, *107*, 135–152.

Chiappe, P. Siegel, L. S., & Gottardo, A. (2002). Reading-related skills in kindergartners from diverse linguistic backgrounds. *Applied Psycholinguistics*, *23*, 95–116

Chiappe, P., Siegel, L., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading*, *6*, 369–400.

Cisero, C. A., & Royer, J. M. (1995). The development and cross-language transfer of phonological awareness. *Contemporary Educational Psychology*, *20*, 275–303

Dutro, S. & Moran, C. (2002), Rethinking English language instruction: An architectural approach. In G. Garcia (Ed.). *English Learners Reading at the Highest Level of English Literacy*. Newark, DE: International Reading Association.

Engelmann, S., & Carnine, D. (1982). Theory of instruction. New York, NY: Irvington.

Echeverria, J., Vogt, M. E., & Short, D. (2004). *Making Content Comprehensible for English learners: The SIOP model*. 2nd Edition. Boston, MA: Pearson/Allyn & Bacon.

Feldman, K., & Kinsella, K. (2005). Create an active participation classroom. *The CORE Reading Expert Newsletter*. /Newsletters/CORE%202005%20Spring%20Newsletter.pdf

Fitzgerald, J. (1995). English-as-a-second-language learner's cognitive reading: A review of research in the united states. *Review of Educational Research*, *65*, 145–190.

Fitzpatrick, A. R., Julian, M. W., Hauck, M. C., & Dawber, T. E. (2006). *The dimensionality of two NCLB tests designed to assess students' English language proficiency*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Genesee, F., Lindholm-Leary, K., Saunders,W., & Christian, D. (2006). *Educating English Language Learners: A synthesis of research evidence*. New York: Cambridge University Press.

Gersten, R., & Woodward, J. (1995). A longitudinal study of transitional and immersion bilingual education programs in one district. *Elementary School Journal*, *95*, 223–240.

Geva, E., Wade-Woolley, L., & Shany, M. (1993). The concurrent development of spelling and decoding in two different orthographies. *Journal of Reading Behavior, 25*, 383–406.

Geva, E., & Yaghoub-Zadeh, Z. (2006). Reading efficiency in native English-speaking and English-as-a-second-language children: The role of oral proficiency and underlying cognitive-linguistic processes. *Scientific Studies of Reading*, *10*, 31–57

Geva, E., Yaghoub-Zadeh, Z., & Schuster, B. (2000). Part IV: Reading and foreign language learning: Understanding individual differences in word recognition skills of ESL children. *Annals of Dyslexia*, *50*, 121–154.

Gottardo, A. (2002). The relationship between language and reading skills in bilingual Spanish-English speakers. *Topics in Language Disorders*, *22*, 46–70.

Haager, D., & Windmueller, M. (2001). Early literacy intervention for English language learners at-risk for learning disabilities: Student outcomes in an urban school. *Learning Disability Quarterly*, *24*, 235–250.

Lafrance, A., & Gottardo, A. (2005). A longitudinal study of phonological processing skills and reading in bilingual children. *Applied Psycholinguistics*, *26*, 559–578.

Leafstedt, J. M., Richards, C. R., & Gerber, M. M. (2004). Effectiveness of explicit phonological-awareness instruction for at-risk English learners. *Learning Disabilities: Research & Practice*, *19*, 251–161.

Lesaux, N. K., and L. S. Siegel. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology*, *39* (6), 1005–1019.

Lesaux, N. K., Lipka, O., & Siegel, L. S. (2006). Investigating cognitive and linguistic abilities that influence the reading comprehension skills of children from diverse linguistic backgrounds. *Reading and Writing: An Interdisciplinary Journal*, *19*, 99–131.

Lesaux, N., & Siegel, L. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology*, *39*, 1005–1019.

Limbos, M. (2006). Early identification of second language students at risk for reading disability. *Dissertation Abstracts International Section A: Humanities and Social Sciences*. 66 (10-A), 2006, pp. 3566.

Limbos, M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities*, *34*, 136–151.

Manis, F.R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K–2 in Spanish-speaking English language learners. *Learning Disabilities Research & Practice*, *19*, 214–224.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice*, *21*, 30–43.

Morrison Institute for Public Policy (2006). Why Some schools with Latino children beat the odds and others don't. Morrison Institute for Public Policy, School of Public Affairs, College of Public Programs, Arizona State University: Tempe, AZ.

Nagy, W. E. (1988). *Teaching Vocabulary to Improve Reading Comprehension*. Neward, DE: International Reading Association.

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

Proctor, C. P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology*, *97*, 246–256.

Quiroga, T., Lemos-Britton, Z., Mostafapour, E.,Abbott, R. D., & Berninger, V. W. (2002). Phonological awareness and beginning reading in Spanish-speaking ESL first graders: Research into practice. *Journal of School Psychology*, *40*, 85–111.

Ramirez, J. D., Yuen, S. D., Ramey, D. R., & Pasta, D. (1991). *Final Report: Longitudinal study of immersion strategy, early-exit, and late-exit transitional bilingual education programs for language minority children*. Volume 1 and 2. San Mateo, CA: Aguirre International, Inc.

Report from the National Reading Panel. (2000). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. Bethesda, Md.: National Institute of Child Health and Human Development.

Rousseau, M. K., Tam, B. K. Y., Ramnarain, R. (1993). Increasing reading proficiency of language-minority students with speech and language impairments. *Education and Treatment of Children*, *16*, 254–271.

Samuels, S.; N. Schermer; and D. Reinking. (1992). Reading fluency: techniques for making decoding automatic. In *What Research Has to Say About Reading Instruction*. Edited by S.J. Samuels, J. Samuels, and A. E. Farstrup. Newark, Del.: International Reading Association,124–143.

Saunders, W. M., Foorman, B. P., & Carlson, C. D. (2006). *Do we need a separate block of time for oral English language development in programs for English learners?*

Schatschneider, C., Carlson, C. D., Francis, D. J., Foorman, B. R., & Fletcher, J. M. (2002). Relationship of rapid automatized naming and phonological awareness in early reading development: Implications for the double-deficit hypothesis. *Journal of Learning Disabilities*, *35*, 245–256.

Scarcella, R. (2003). *Academic English: A conceptual report*. Technical Report 2003-1. The University of California Linguistic Minority Research Institute Technical Report. Santa Barbara, CA, University of California, Santa Barbara.

Shanahan, T., & August, D. (2006). *Report of the national literacy panel: Research on teaching reading to English language learners*. Mahwah, NJ: Lawrence Erlbaum.

Skindrud, K., & Gersten, R. (2006). An independent evaluation of two prominent reading reforms in the Sacramento City school: Academic and special education outcomes. *Elementary School Journal*.

RETURN to TABLE of CONTENTS

Snow, C. E., and Wong-Fillmore, L. (2000). *What Teachers Need to Know about Language*. ERIC Clearinghouse on Language and Linguistics Special Report. http://www.cal.org/resources/teachers/teachers.pdf

Snow, M.A., Met, M., & Genesee, F. (1989). A conceptual framework for the integration of language and content in second/foreign language instruction. *TESOL Quarterly*, *23*(2), 201–218.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Swanson, H. L., Sáez, L., & Gerber, M. (2004). Do phonological and executive processes in English learners at risk for reading disabilities in grade 1 predict performance in grade 2? *Learning Disabilities Research & Practice*, *19*, 225–238.

Verhoeven, L. (1990). Acquisition of reading in a second language. *Reading Research Quarterly*, *25*, 90–114.

Verhoeven, L. T. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading*, *4*, 313–330.

Wade-Woolley, L., & Siegel, L. S. (1997). The spelling performance of ESL and native speakers of English as a function of reading skill. *Reading and Writing: An Interdisciplinary Journal*, *9*, 387–406.

Wang, M. & Geva, E. (2003). Spelling acquisition of novel English phonemes in Chinese children. *Reading and Writing: An Interdisciplinary Journal*, *16*, 325–348.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education*, *26*, 207–214.

Wong, S. D., Fillmore, L., & Snow, C. (2002). What teachers need to know about language. In C.T. Adger, C. E. Snow, & D. Christian (Eds.), *What Teachers Need to Know about Language* (pp. 7–53). McHenry, IL: Delta Systems and Center for Applied Linguistics.

RETURN to TABLE of CONTENTS

Chapter 6

Assessing Comprehension and Communication in English State to State for English Language Learners (ACCESS for ELLs®)

Jim Bauman, Tim Boals, Elizabeth Cranley, Margo Gottlieb, and Dorry Kenyon

In 2002, two consultants at the Wisconsin Department of Public Instruction and a consultant from the Illinois Resource Center outlined a plan for promoting a partnership of states that would support the development of a system based on English language proficiency standards and assessments. The standards and assessments would be aligned to specific content area language to facilitate English language learners' access to language arts, mathematics, science, and social studies. The plan was designed to meet not only the requirements of the new No Child Left Behind (2001) legislation but also the recommendations of research in language education that, for the last two decades, has called for teaching language through academic content.

The outline became the foundation of a federal enhanced assessment grant proposal that included an initial partnership of three relatively small states, Wisconsin, Delaware, and Arkansas. In addition to these states, the World-Class Instructional Design and Assessment (WIDA) Consortium included key partnerships with the Center for Applied Linguistics (CAL), the University of Wisconsin System, the University of Illinois, Urbana-Champaign, and several experts in the education of English language learners and language testing.

Within two months of receiving funding, in March 2003, the District of Columbia, Maine, New Hampshire, Rhode Island, and Vermont contacted WIDA and made the decision to join this new cooperative of states that aspired to share resources and work together to improve teaching, learning, and assessment for English language learners. By the end of that year, Illinois had joined the consortium, followed soon thereafter by Alabama.

During 2003, the Consortium drafted the *WIDA English Language Proficiency Standards for English Language Learners in Kindergarten through Grade 12* (State of Wisconsin, 2004) and began work on *Assessing Comprehension and Communication in English State to State for English language learners* (ACCESS for ELLs®) English language proficiency test. WIDA pilot and field tested the new assessment in 2004, and by spring 2005, the test was operational in three states: Alabama, Maine, and Vermont. In spring 2006, twelve WIDA Consortium member states tested students

RETURN to TABLE of CONTENTS

using ACCESS for ELLs®. At that time, World-Class Instructional Design and Assessment (WIDA) had moved the central office from the Wisconsin Department of Public Instruction to the University of Wisconsin–Madison's Wisconsin Center for Education Research (WCER).

WIDA's aligned assessment system begins with the theory and research that has informed the development of its English language proficiency standards (hereafter referred to as the WIDA Standards). The WIDA Standards are the grounding and anchor for test specifications which delineate the parameters for the development of test items. In the sections that follow, we briefly discuss these theoretical foundations and examine multiple aspects of ACCESS for ELLs® and its development process.

The Theoretical Basis for ACCESS for ELLs®

The theoretical underpinnings for WIDA's English language proficiency test are drawn from the fields of second language acquisition research and linguistics. The seminal work of Cummins (1981), validated by Lindholm-Leary (2001) and Collier & Thomas (2002), offers a continuum of expected performance of English language learners as they progress through the language development process to acquire the English language skills necessary for reaching academic parity with their English proficient peers. Coupled with the exploration of the construct of academic language proficiency by Bailey & Butler (2002) and Scarcella (2003), among others, WIDA developed a model of academic language proficiency (Gottlieb, 2003) to guide the formulation of the WIDA Standards and their accompanying model performance indicators (Gottlieb, 2004). Furthermore, linguistic (e.g., vocabulary usage, language complexity, phonological and syntactic development) and pragmatic (e.g., functional language) aspects of communicative competence are recognized as criteria that delineate the performance definitions for WIDA's five levels of English language proficiency, identified in the WIDA Standards, listed in Table 1.

ACCESS for ELLs®, an operational representation of the standards, reflects an identical philosophical and theoretical basis. The connections between the theoretical basis of communicative competence and that of second language teaching and testing, first recognized by Canale and Swain (1980), are captured in WIDA's English language proficiency standards and test.

The Use of Standards as a Base for Test Development

The conceptualization and formulation of the WIDA Standards have been central to the design of ACCESS for ELLs®. In 2003, a group of sixty educators from eight member states met to begin the development process. WIDA wanted to ensure two essential elements: 1) a strong representation of the *language* of state academic standards across the core content areas; and 2) consensus by member states on the components of the English language proficiency standards. As new states have joined the Consortium, teams of researchers have continued the process by systematically conducting alignment studies between the WIDA Standards and a new state's content standards. In the initial forms of ACCESS for ELLs®, there has been a 1:1 correspondence of test items to the performance indicators for each standard in WIDA's large-scale framework. With this clear match, construct validation has been built into the test.

Although the WIDA Standards have remained constant during the past two years, WIDA has augmented the

Table 1. WIDA English Language Proficiency Standards

Standard	Description
1	English language learners communicate in English for social and instructional purposes in the school setting.
2	English language learners communicate information, ideas and concepts necessary for academic success in the content area of language arts.
3	English language learners communicate information, ideas and concepts necessary for academic success in the content area of mathematics.
4	English language learners communicate information, ideas and concepts necessary for academic success in the content area of science.
5	English language learners communicate information, ideas and concepts necessary for academic success in the content area of social studies.

RETURN to TABLE of CONTENTS

strands of model performance indicators (PIs). Building on the WIDA Standards, Teachers of English to Speakers of Other Languages (TESOL) published its *PreK–12 English language proficiency standards* in 2006 (Gottlieb, Carnuccio, Ernst-Slavit, & Katz). In working on English language proficiency standards from state and national perspectives, WIDA has come to realize the necessity of having strands of performance indicators that are flexible and dynamic rather than static. The strong tie between the standards and the test will remain as this principle is made operational in the next iterations of ACCESS for ELLs®.

Comparison between the New and Prior Generation of ELP Tests

The language proficiency tests of the prior generation were generally constructed in response to legislation and litigation of the 1970s, starting with the Lau v. Nichols Supreme Court decision in 1974. They fulfilled a need at a time when very few instruments were available to assess the language proficiency of linguistically and culturally diverse students in the United States. As such, in large part, they represented the thinking of behavioral and structural linguistics prevalent during the 1960s. In contrast, ACCESS for ELLs® was born from the need for an enhanced assessment of English language proficiency to comply with the No Child Left Behind Act. WIDA strongly felt that ACCESS for ELLs® needed to fulfill those requirements as well as reflect current theory, research, and best educational practice for English language learners.

Table 2 compares the features of the prior generation of English language proficiency tests and ACCESS for ELLs®. These differences not only reflect the shift in theory, research, and educational practice in the last decades but, significantly, the change of purpose to one of accountability.

Creating the Test Blueprint

Early on, the WIDA Consortium recognized that two competing forces had to be reconciled in creating a successful test based on the WIDA Standards. The first of these was the need to comprehensively assess performance indicators across the full set of ELP standards in four language domains and five levels of English language proficiency. The second was the need to keep the total test time per student and the total test administration burden per school or district within acceptable limits. The solution framed itself around two reasonable expectations: (1) that language proficiency develops apace, within, and across the four domains, and (2) that a student only needs to be tested on the subset of items that straddle his or her true language proficiency.

Table 2. Comparison of ACCESS for ELLs® with other English Language Proficiency Tests

Prior Generation of Tests	ACCESS for ELLs®
Not based on standards	Anchored in WIDA's English language proficiency standards
Non-secure, off-the-shelf, low stakes test	Secure; high stakes test
Social language emphasized	Academic language emphasized
Not aligned with academic content standards	Aligned/linked with core academic content standards
In general, integrated oral language domains	Independent oral language domains (i.e., listening and speaking)
Different tests used for each grade level cluster (no comparability)	Vertically scaled across grade level clusters
One test used for each grade level cluster	Divided into tiers within each grade level cluster to accommodate a range of contiguous proficiency levels
Non-compliant with No Child Left Behind	Compliant with No Child Left Behind
Identical test used for screening, identification, placement, and reclassification	ACCESS for ELLs® used for annual assessment, and WIDA-ACCESS Placement Test (W-APT) used for screening and identification
Static with irregular updates	Dynamic with updates and improvements every year documented in its annual Technical Report

RETURN to TABLE of CONTENTS

The test blueprint realized these expectations through a tiered design. Three overlapping forms of the test were designed, one mapped to proficiency levels 1, 2, and 3; a second to levels 2, 3, and 4; and a third to levels 3, 4, and 5. An individual student, broadly identified as falling within one of these three tier ranges, would be given only that tier of the test. This convention cut the overall test administration time per student to approximately 2.25 hours.

The test blueprint also needed to consider the demands of academic content and how it differs and develops across grades. Since the WIDA Standards specifically address language proficiency in academic settings, test items needed to address, though decidedly not test, academic content. The most equitable solution—that of developing specific tests at each grade level—was rejected as impractical, not just because of cost, but also because of the much higher demands it would make for test administration time. The compromise here was to group test organization into grade level clusters, initially K–2, 3–5, 6–8, and 9–12, following the layout of the performance indicators in the WIDA Standards. Later, kindergarten was split out into its own "cluster," in recognition of the large developmental leap taking place between kindergarten and first grade.

Within all grade level clusters, except kindergarten, the same item types and test administration procedures are used. The listening and reading components of tests, reflecting their receptive nature, lend themselves to group administration using a traditional multiple-choice item format. The speaking and writing components, reflecting their productive nature, are realized in performance-based tasks and scored against their respective rubrics. In the speaking test, these tasks are presented in an interview or question-answer format and are administered individually and adaptively. In the writing test, they are presented in a short-answer or essay format and are administered in a group setting.

Developing Test Items

The creation of items for ACCESS for ELLs® begins with a formal specification for each item's properties. The specification directly addresses one or more performance indicators (PI) from the WIDA Standards document, in particular, those PIs appropriate for large scale (summative) test objectives. Because the test items are written according to the Model Performance Indicators of the WIDA Standards, they reflect in their coverage and form the academic language requirements specified in the standards.

Individual ACCESS for ELLs® test items are embedded in the context of a content-based theme, and that theme incorporates items at different proficiency levels. Therefore, the specification is first aimed at describing the theme components—orientation and theme stimulus—and then, detailing the item characteristics. These characteristics include the item's proficiency level, the graphic and linguistic structure of the item stimulus, the linguistic properties of the task statement (formulated as one or more questions, prompts, or models), and the expectations of the response. For multiple-choice items on the listening and reading tests, the response options may include text, graphics, or both. For writing and speaking items, the response expectations are specified in scoring rubrics designed for each of these two domains.

The theme stimulus always has some graphic elements and may also include text elements. The relative weight given to one or other of these two types of stimuli depends on the language domain being tested as well as the content standard. Graphics are intended to reduce the potentially confounding influence of whatever linguistic channel is used to present the task context by opening a visual channel to frame that context. From another vantage point, the graphics also provide a non-linguistic means of supplying English language learners with necessary background knowledge to compensate for the advantage that students with academic preparation might otherwise have. The net effect of the use of theme graphics, then, is to increase the redundancy of task-specific contextual information. A concomitant effect is, typically, that the student test taker will have multiple pathways to finding or producing a correct or appropriate response. This notion ties importantly to our contention that ACCESS for ELLs® does not test individual skills or mechanical processing abilities, but tests language proficiency in a more comprehensive sense.

The great majority of test items are initiated by teachers in the WIDA Consortium states. The consortium "assembles" the teachers through an online item writing course offered annually. During the course, the teachers learn the underpinnings of the test framework, item writing, and test specifications. Teachers write items in all four language domains. Items undergo two formal in-house reviews and edits, the first to prepare items for a content and bias review session with a second set of educators, and the second to finalize item content. The first review focuses

RETURN to TABLE of CONTENTS

on assuring that each item meets the requirements of its PI(s) and that the graphic and text components of the item are well integrated. The second review session incorporates suggestions of the content and bias reviewers and strives for appropriate distracter balance in the reading and listening items.

Ultimately, the Rasch measurement model, upon which the test is built, is used for the empirical analysis of the quality of test items following piloting and field testing. At the item level, Rasch mean square infit and outfit statistics are examined to ensure only items that fit the Rasch measurement model (i.e., measure the same construct) are included in operational test forms. On an annual basis, each item's empirical item difficulty is also compared against its target proficiency level to help us better understand the characteristics of the test items vis-à-vis the proficiency levels defined in the WIDA Standards. This analysis helps us refine the item specifications and the item review procedure for subsequent operational forms. In addition, items are examined for differential item functioning (male versus female, speakers of Spanish as a home language versus all other English language learners) to flag items that may need to be replaced or revised.

Creating the Operational Forms

Following the creation of the item pool, thematic folders of test items are selected for inclusion in the test booklets. This selection process considers the spread of content topics, balance in cultural representation, and fit to the measurement model.

Within each reading and listening test booklet, at least one thematic folder is included for each of the five WIDA Standards. Typically, two or three additional folders are added to bring the total item count to between 25 and 30 in order to reach acceptable levels of reliability and discrimination ability. These numbers were determined to be adequate through field testing.

The speaking test incorporates three folders, addressing the following standards:

- Standard 1 (social and instructional language) in the first folder with tasks at proficiency levels 1, 2, and 3;
- Standards 2 and 5 (the language of language arts and of social studies) in the second folder with tasks at all five proficiency levels; and
- Standards 3 and 4 (the language of mathematics and of science), in the third folder with tasks at all five proficiency levels.

The folders are administered in the same sequence. Within each folder, the test administrator presents tasks at the lowest proficiency level first and continues with higher level tasks until the student reaches his or her performance ceiling, at which point the next thematic folder is introduced and administered in the same way. Because the speaking test is adaptive and individually administered, it is not divided into tiers.

The writing test is organized around three thematic folders for the first tier and four folders each for the second and third tiers. The following table summarizes the organization of the folders by tier.

The fourth writing test folder for the two higher tiers is referred to as an "integrated task." In scope, it is the most challenging of the tasks and is aimed at the highest language proficiency levels. Where the first three tasks are intended to take approximately 10 to 15 minutes each, the integrated task is intended to take about 30 minutes. Given this scope, it incorporates three language proficiency standards: social and instructional, language arts, and social studies.

The kindergarten test differs from all the other tests in that it is administered in its entirety individually, and all components are adaptive. Again, because of this design, there are no tiers. In other respects, only the kindergarten writing test differs substantially in substance and structure

Table 3. Structure of the ACCESS for ELLs® Writing Tests

	Tier A	Tier B	Tier C
Folder 1	SI (Level 3)	SI (Level 4)	SI (Level 4)
Folder 2	MA (Level 3)	MA (Level 4)	MA (Level 5)
Folder 3	SC (Level 3)	SC (Level 4)	SC (Level 5)
Folder 4		IT (SI, LA, SS) (Level 5)	IT (SI, LA, SS) (Level 5)

Note: "Level" represents the highest proficiency level aimed at by the task

RETURN to TABLE of CONTENTS

from the other test components: Its thematic folders reflect the need to incorporate only the three lowest levels of the standards for kindergarten content. Consequently, no tasks beyond proficiency level 3 are included. In future versions of ACCESS for ELLs®, the kindergarten test will be developed around its own set of performance indicators, rather than being built on the K–2 cluster.

Pilot and Field Testing

Pilot testing for ACCESS for ELLs® took place in two rounds in 2004 with 1,244 students in five representative districts across three participating WIDA states. The results of the pilot confirmed that the overall test format and the concept of tiers were appropriate, as was the test content for the different grade level clusters. Furthermore, the prototypical items chosen to assess the different proficiency levels were adequate. Finally, the pilot test allowed us to refine the test administration instruction and to determine administration times for the various sections of the test.

After the completion of the pilot test, development of the item and task pool for the field test of two forms of ACCESS for ELLs® was completed, followed by a content and bias review with educators from seven WIDA states. Field testing of the two forms took place in eight WIDA states with approximately 7,000 students sampled proportionately from each state (about 6% of the ELLs in each state). State coordinators facilitated the participation of a wide variety of districts within each state. The results of the field test—the analysis of which focused most closely on item characteristics—confirmed the construct of the test and provided data for equating and scaling that would create usable scores.

Standard Setting Process

In order to interpret what ACCESS for ELLs® scores mean, standard-setting studies were conducted in Madison, WI, between April 20 and 27, 2005. The purpose of the studies was not to "set" new standards on WIDA's ACCESS for ELLs® per se, but to conduct a defensible and replicable approach to determining the relationship between student performance on the four domains of ACCESS for ELLs® and the proficiency levels defined by the WIDA Standards. This was done using the WIDA Standards together with empirical information from the field test data. The following is a brief summary of the standard setting study; for a fuller description, see ACCESS for ELLs® Technical Report 1, *Development and Field Test of ACCESS for ELLs®*.

Four panels of 20 to 22 teachers and administrators were convened, one for each major grade-level cluster: 1–2, 3–5, 6–8, and 9–12. For the listening and reading assessment, a bookmarking procedure (Mizel, Lewis, Patz & Green, 2001) was used. Panelists were given books with all items within their grade cluster arranged from least difficult to most difficult (based on the empirical data). After discussing the pertinent performance indicators (PIs) and performance level definitions from the WIDA Standards, panelists read through the items and placed a bookmark at the item that they felt a student at proficiency level 1 would have a 50% chance of answering correctly. They then repeated this procedure for all levels up to the level 5/6 border. During this procedure, panelists worked independently, followed by an opportunity to discuss the results, reconsider, and, if they chose, adjust their bookmarking. The final results, based on the average item difficulty across all panelists, were compiled and presented to the WIDA management team, who used these data to help determine the final cut scores.

For writing and speaking assessments, modifications of the body-of-work method (Kingston, Kahl, Sweeney & Bay, 2001) were used. In the modification used for the writing assessment, portfolios were presented to the panelists in order of raw score. For writing, the panelists were presented a book containing between 17 and 22 portfolios of student responses from their grade cluster. Each portfolio consisted of a student's responses to the four writing tasks. Portfolios were presented in ascending order; the first portfolio represented a student's work that had received the lowest total raw score across the four pieces of writing and the last portfolio presented was that of a student with the highest total raw score on the four pieces of writing. We attempted to present portfolios at equal intervals from lowest to highest; that is, each succeeding portfolio had been scored three to four raw points higher than the preceding one. As with the bookmarking procedure, panelists began by discussing as a group the pertinent PIs and performance level definitions from the WIDA Standards. Then, they read the portfolios one at a time. Working individually, each panelist made a judgment as to the probability that the samples of work in the portfolio represented the writing of a student at a given WIDA proficiency level. After the panelists made their judgments for a portfolio, the results were compiled and the panelists discussed them as a group, followed by

RETURN to TABLE of CONTENTS

the opportunity to reconsider and adjust their judgments if they so chose., The average for each portfolio in each category across all panelists, was used as input into a logistic progression procedure to determine the points along the underlying writing proficiency continuum at which at least 50% of the panelists would be expected to agree that the writing represented the work of the next higher proficiency level rather than the current proficiency level. The results from this analysis were used to set the cut scores for the proficiency levels.

The modified body-of-work procedure for the speaking assessment was similar. The panelists listened to portfolios of students responding to speaking tasks across the entire speaking test administered to them. After each portfolio, the panelists recorded their judgments and then discussed the results.

Panelists evaluated the materials and processes used in each standard setting study, as well as how confident they were in the cut scores that they had set using the process. While the panelists showed a great deal of confidence for all procedures, the process for the writing assessment appeared to be met with the greatest satisfaction.

Validation Process

Several initial studies were conducted by the WIDA Consortium that support the use of the test as an assessment of English language proficiency. The first group of studies examined the relationship between performance on ACCESS for ELLs® and external criteria (criterion-related) validity. The second group of studies examined the performance of the ACCESS for ELLs® test items on the basis of internal criteria (content) validity.

The reliability of test scores is a necessary (although not sufficient) requirement for test validity. For most users of ACCESS for ELLs®, decisions about student English language proficiency are based on the overall composite score. Results from the technical analysis of the first operational administration of ACCESS for ELLs® (Series 100) indicate that the reliability (using a stratified Cronbach alpha coefficient) of the overall composite score is very high across all grade-level clusters. For kindergarten, the coefficient was .930; for grades 1–2, .949; for grades 3–5, .941; for grades 6–8, .933; and for grades 9–12, .936. Using the approach of Livingston and Lewis (1995) and Young and Yoon (1998) to investigate the accuracy and consistency of classification, the accuracy of decisions about student placement around the cut score of proficiency levels 5 (bridging) and 6 (reaching) using the composite score was likewise very high across all grade level clusters: .975 for grades 1–2; .972 for grades 3–5; .976 for grades 6–8; and .977 for grades 9–12. (Students in kindergarten cannot achieve a composite score at level 6.) The accuracy of decisions at the cut score between levels 2 (entering) and 3 (developing) were also high: .949 for kindergarten; .943 for grades 1–2; .940 for grades 3–5; .936 for grades 6–8; and .921 for grades 9–12.

With reliability established, two main research studies examined performance on ACCESS for ELLs® against external criteria. In the first, using data collected during the field test in the fall of 2004, performance on ACCESS for ELLs® had a moderate-to-strong relationship to students' proficiency level designation according to their local (district) records. These designations were based on policies and procedures in place prior to the introduction of ACCESS for ELLs®, and varied greatly across the eight states and many districts participating in the field test. Across the WIDA Consortium, at least four main English proficiency tests were used to help make these designations and some districts used only four levels (plus "monitored") to designate English language learners (rather than the five used by WIDA). Nevertheless, within districts and states, these levels are hierarchical, and it was expected that, as a group, students designated as having higher proficiency levels (prior to participating in the ACCESS for ELLs® field test) a priori should have more English language proficiency than students at lower levels.

Results from this research indicated that across all four domains (listening, speaking, reading and writing) and across all grade level clusters (from 1 to 12), the average ACCESS for ELLs® scale score obtained by students in the field test—according to their a priori proficiency-level assignment—increased as their a priori proficiency level designation increased (their a priori level being an assessment of student proficiency based on criteria external to the WIDA Standards and ACCESS for ELLs® assessment system). Rank order correlations between the a priori proficiency level designations and ACCESS for ELLs®domain scores ranged from .282 (listening in grades 1–2) to .698 (speaking in grades 9–12).

The second main study that related performance on ACCESS for ELLs® to external criteria came from the ACCESS for ELLs® bridge study. In this study, about 5,000 students took ACCESS for ELLs® and one of four other

RETURN to TABLE of CONTENTS

English language proficiency tests. Further information on this study and its results are presented in the following section.

As for the content of ACCESS for ELLs®, there is a strong match between the assessment and the standards upon which they are based. The practical definition of the WIDA Standards is embodied through the performance level definitions and the over 400 model performance indicators upon which every item and assessment task in the test is based. In other words, every test item is designed to offer students the opportunity to demonstrate meeting the model performance indicators for their grade-level cluster, in a certain domain, and at a certain proficiency level. These definitions are central to the specifications used to develop and review items and assessment tasks. Panelists in the WIDA ACCESS for ELLs® standard-setting study (see above) made full use of both the performance level definitions and the model performance indicators in setting the standards. As every test form is developed, members of the content review committee make sure items match their model performance indicators. The relationship between the assessment items and tasks and the model performance indicators of the WIDA Standards is intended to be so transparent that sample assessment items prepared for wide distribution to educators in WIDA consortium states are presented along with the targeted model performance indicators.

As a way to investigate the connection between the WIDA Standards and the items empirically, we conducted a study to investigate whether the items embody the five proficiency levels defined by the standards. This study focused on the following three questions:

1. Are the items in ACCESS for ELLs® empirically ordered by difficulty as predicted by the WIDA Standards?
2. Does that ordering differ by domain (listening or reading)?
3. Does that ordering differ by type of standard (language arts, mathematics, etc.)?

Data came from the final calibration of the fall 2004 ACCESS for ELLs® field test forms in listening and reading. The field test included more than 6,500 students in grades 1 to 12 from eight of the WIDA Consortium states. The total number of students tested in the field test represented approximately 3.5% of the ELL enrollment in WIDA Consortium states in fall 2004.

Following the field test, the listening and reading items were vertically scaled (separately) using common item equating in concurrent calibration using Winsteps software, an application of the Rasch measurement model. Items that misfit (and thus were revised or discarded for operational Series 100 of ACCESS for ELLs®) were eliminated from this study. For the different analyses, average item difficulty was calculated. The study found that the average item difficulty indeed increased as the item's a priori proficiency level (i.e., the proficiency level that the item was intended to target) increased. When the study broke out the data by domain (listening and reading), the result was the same with the exception of levels 4 and 5 in reading, which were at about the same difficulty level. When the study broke out the data by standard, the results were also the same, with the exception of level 5 in language arts. As might have been predicted, items assessing social and instructional language were easiest overall; followed by items assessing the language needed for academic success in English language arts. The average item difficulty for the remaining three standards all clustered together.

This investigation into item difficulty on the field test forms of ACCESS for ELLs® provides additional evidence for the claim that the test items assess five English proficiency levels needed for academic success as posited by the WIDA Standards. While the analyses identified a few areas needing further refinement, this study illustrates the type of continued research that will be conducted annually to ensure that the assessment operationalizes the WIDA Standards and that users know that results are valid interpretations of performance based on these standards.

Bridge Study

In order to help WIDA Consortium states understand performance on the English language proficiency tests they had been using prior to ACCESS for ELLs® and performance on ACCESS, the Consortium undertook a bridge study. In this study, 4,985 students enrolled in grades K through 12 from selected districts in Illinois and Rhode Island took ACCESS for ELLs® and one of four older English language proficiency tests: *Language Assessment Scales* (LAS), the *IDEA Proficiency Test* (IPT), the *Language Proficiency Test Series* (LPTS), and the *Revised Maculaitis II* (MAC II). They took the tests in the spring of 2005 within a window of six weeks. The bridge study was concurrent with the first operational administration of ACCESS for

RETURN to TABLE of CONTENTS

ELLs® in Alabama, Maine and Vermont. In every case, the older test was administered first. Older English language proficiency tests were scored following district practice, and ACCESS for ELLs® was centrally scored.

Table 4. Average Correlations

Test	List	Speak	Read	Write
IPT	0.614	0.627	0.658	0.629
LAS	0.514	0.570	0.643	0.561
LPTS	0.610	0.664	0.765	0.707
MAC II	0.468	0.508	0.582	0.545

Note: All levels of each test included within domain.

The data collected from performance on an older test was used to predict performance on ACCESS for ELLs®. A linear regression procedure was used. The data was analyzed at CAL, and the results were made available to the states in the fall of 2005. Many states used the analyses to help them develop Annual Measurable Achievement Objectives (AMAOs) to meet federal reporting requirements.

The study also allowed us to investigate the strength of the relationship between the new ACCESS for ELLs® test and four tests of English language proficiency—a criterion-related validity question. As all five tests claim to measure developing English language proficiency, we expected significant correlations between student performance on ACCESS for ELLs® and the other tests of English language proficiency. On the other hand, ACCESS for ELLs® was developed with a different intent; that is, to assess the English proficiency needed to succeed academically in U.S. classrooms based on clearly defined English language proficiency standards. Because of this more specific scope, we did not expect the correlations to be very strong. If the correlations were very high, one or more of the older tests and ACCESS for ELLs® could be seen as interchangeable, and an argument could be made against introducing a new English language proficiency test. Thus, moderate correlations between scores in listening, speaking, reading, and writing on ACCESS for ELLs® and scores in those domains on other tests were expected. Low correlations, however, would have been very troubling, possibly indicating that the two tests were measuring very different constructs.

Overall, the study found moderate-to-high correlations, as predicted, between the various forms of the older generation tests across the grade-level clusters and ACCESS for ELLs®. Table 4 presents the average correlation, by language domain, found between scores on the different forms (levels) of each older-generation test and performance on ACCESS for ELLs®.

Table 5 presents, for each ACCESS for ELLs® domain and by test, the lowest and the highest correlation found across all the forms (levels) available for any given test in that domain. Across the four older-generation tests, across their separate and non-interchangeable levels, and across the four language domains assessed by ACCESS for ELLs®, there was a consistent finding of a moderate-to-strong correlation between student performance on ACCESS for ELLs® and on the older tests. This result provides strong initial support to the claim that performance on ACCESS for ELLs® represents an assessment of English language proficiency, just as the older-generation tests claimed. However, the absence of very high correlations provides some support to the claim that the standards-based, NCLB-compliant ACCESS for ELLs® is assessing the construct of English language proficiency somewhat differently and is *not* interchangeable with the older-generation tests.

Additionally, the data allowed us to investigate the relationship between the cut scores on the older tests and the level 5/6 cut score on ACCESS for ELLs®. Although there was variation across tests, language domains, and grade level clusters, in every case, the level 5/6 cut on ACCESS for ELLs® was higher than the score recommended as the "exit cut" on the other tests, as predicted by the ACCESS for ELLs® bridge study results. This result may also provide support that the type of English proficiency assessed by

Table 5. Range of Correlations (Low and High) By Test and ACCESS for ELLs® Domain

	List		Speak		Read		Write	
Test	Low	High	Low	High	Low	High	Low	High
IPT	0.515	0.712	0.594	0.767	0.540	0.741	0.550	0.776
LAS	0.474	0.525	0.548	0.599	0.317	0.757	0.323	0.684
LPTS	0.532	0.666	0.600	0.695	0.658	0.822	0.529	0.759
MAC II	0.300	0.599	0.330	0.621	0.362	0.675	0.175	0.685

RETURN to TABLE of CONTENTS

ACCESS for ELLs® is more demanding than the older generation tests, which, for the most part, were not developed to assess academic language.

For further information about any of the studies reported here, or additional studies to date regarding the validity of ACCESS for ELLs®, see the following reports: Kenyon (2006), Technical Report 1, *Development and Field Test of ACCESS for ELLs®*; Gottlieb & Kenyon (2006), *The WIDA Bridge Study* Technical Report 2, and Kenyon et al (2006), *Annual Technical Report for ACCESS for ELLs®, Series 100, 2005 Administration*, Annual Technical Report No. 1). Further validation studies on ACCESS for ELLs® are currently underway based on our research agenda crafted with input from WIDA member states.

Scoring and Reporting on ACCESS for ELLs®

ACCESS for ELLs® is scored and reported by MetriTech, Inc., who sends to each participating district: (1) reports for parents and guardians for which scores are reported as simple bar graphs in a multitude of languages, (2) a comprehensive report aimed at teachers and program coordinators at the school level, and (3) frequency reports for districts and schools.

Score reports for the test include vertically scaled scores across the K–12 grade span (100–600 points) in each domain (listening, speaking, reading, and writing), and an interpretation of the scale score designated as an English proficiency level from 1.0 to 6.0. Reports include four weighted composite proficiency scores:

- an Overall composite score reflecting all domains,
- an Oral Language composite score (listening and speaking),
- a Literacy composite score (reading and writing), and
- a Comprehension composite score (listening and reading).

The member states of the WIDA consortium determined the relative weights of the composite scores. The Overall composite score is weighted as 15% listening, 15% speaking, 35% reading, and 35% writing. The Oral and Literacy composite scores are weighted 50% for each domain, and the Comprehension composite score is 30% listening and 70% reading. In addition to the domain and composite scores, teachers are given information in the form of raw scores of how each child performed in relation to the WIDA Standards.

ACCESS for ELLs® scores illustrate where along the continuum of academic language development a student is performing at the time of the test. Scores give teachers, students and families a reference for estimating growth toward proficiency in the language skills necessary to be successful in academic content area classes. Coupled with the WIDA Standards, scores, along with an Interpretative Guide, help teachers better understand their ELL students' language capabilities and needs. ACCESS for ELLs® scores also serve as one criterion for program exit decisions as well as a vital piece of evidence in accountability models.

Conclusion

The WIDA Consortium is a cooperative of states working together to develop and implement standards and assessments that are aligned with best practices for teaching and assessing English language learners. ACCESS for ELLs® is one essential result of this cooperative venture. ACCESS for ELLs® continues to progress through use, development and research, with the understanding that test scores must be sufficiently reliable and valid for determining student progress over time. Reliable and valid test scores are equally important at the local level for making programmatic decisions and tailoring individual student support plans to the needs of English language learners. Over time, accurate and more comprehensive data regarding students' English language development will allow all stakeholders to see growth patterns and assess the extent to which programs serving ELLs are working as intended.

The WIDA Standards, and by extension ACCESS for ELLs®, address language proficiency that research has consistently identified as necessary for ELLs to reach to be succeed in general education classrooms: proficiency in specific academic and technical content at the higher language levels. By aligning ACCESS for ELLs® so closely and transparently to the WIDA Standards and reporting student results by standard, WIDA also encourages teachers to use more innovative instructional strategies that combine the teaching of academic language **and** the teaching of content. Furthermore, this focus on academic language encourages schools to maintain academically aligned curriculum and program support to ensure educational continuity and rigor necessary for student success.

RETURN to TABLE of CONTENTS

References

Bailey, A. L., & Butler, F. A. (2002). *An evidentiary framework for operationalizing academic language for broad application to K–12 education: A design document*. Los Angeles: University of California, Los Angeles, National Center for the Study of Evaluation/ National Center for Research on Evaluation, Standards, and Student Testing.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1–47.

Collier, V. P., & Thomas, W.P. (2002). *A national study of school effectiveness for language minority students' long term academic achievement*. University of California, Santa Cruz: Center for Research on Education, Diversity and Excellence.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education, *Schooling and language minority students: A theoretical framework* (pp. 3–49). Sacramento, CA. California Department of Education.

Gottlieb, M. (2003). *Large-scale assessment of English language learners: Addressing educational accountability in K–12 settings*. TESOL Professional Papers #6. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Gottlieb, M. (2004). Overview. In *WIDA consortium K–12 English language proficiency standards for English language learners: Frameworks for large-scale state and classroom assessment. Overview document*. Madison, WI: State of Wisconsin.

Gottlieb, M., Carnuccio, L., Ernst-Slavit, G., & Katz, A. (2006). *PreK–12 English language proficiency standards*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Kingston, N.M., Kahl, S.R., Sweeny, K.P., and Bay, L. (2001). *Setting performance standards using the body of work method. In G.J. Cizek (Ed.). Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Lindholm-Leary, K. (2001). *Dual language education*. Avon, England. Multilingual Matters.

Livingston, S.A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.

Mitzel, H.C., Lewis, D.M, Patz, R.J., and Green, D.R. (2001). The bookmark procedure: psychological perspectives. In G.J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Scarcella, R. (2003). *Academic English: A conceptual framework*. (Technical Report 2003–1.) Santa Barbara, CA: The University of California Linguistic Minority Research Institute.

Young, M.J., & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* (CSE Tech. Rep. 475). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies.

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

Chapter 7

Overview of Existing English Language Proficiency Tests

Susan G. Porter and Jisel Vega

Over the years, many formal and informal assessments have been developed and used for the purpose of measuring English language proficiency of students whose home language is not English. Many of these assessments, however, do not meet the requirements specified in Title III of the NCLB Act. The law requires that these assessments:

- Measure annual progress in English language proficiency in the four domains of reading, writing, listening, speaking, and a separate measure for comprehension
- Be aligned to state-developed English language proficiency (ELP)/English language development (ELD) standards
- Relate to the state-adopted content standards, including those for reading/language arts and mathematics.

In addition, each state is encouraged to align its ELD standards to the state content standards and assessments.

In order to support the development of compliant English language proficiency tests, the U.S. Department of Education provided funding through the Enhanced Assessment Grant under NCLB, § 6112b. Four consortia received research support atnd four tests were developed: American Institute for Research (AIR) and the State Collaborative on Assessment and Student Standards (LEP-SCASS) collaborated on the English Language Development Assessment (ELDA); AccountabilityWorks in collaboration with several states developed the Comprehensive English Language Learner Assessment (CELLA); A World-Class Instructional Design and Assessment (WIDA) Consortium in collaboration with several states developed Assessing Comprehension and Communication in English State to State for English Language Learners® (ACCESS for ELLs); and Mountain West Assessment Consortium (MWAC) in collaboration with several states and Measured Progress developed the Mountain West Assessment (MWA).

While many states are using an assessment developed by one of the consortia, other states developed their own assessments with the assistance of outside test developers. In several instances, states opted to use commercially available assessments, which are either "off the shelf" or augmented and aligned versions of assessments. Table 1 lists the assessments that states are currently using to meet

DISCLAIMER: *The information presented in this chapter is as accurate and current as possible at time of publication. We welcome changes to the technical data so that it may be incorporated into the web-based version of this report over time. For the web-based version of this report, please see the UC Davis School of Education web site at http://education.ucdavis.edu/research/ELP_Assessment.html.*

RETURN to TABLE of CONTENTS

Title III requirements. The table also indicates the name of the test developer(s) and the date of implementation.

The purpose of this chapter is to provide summary information on language proficiency assessments that states are currently using for Title III purposes. The development history and the technical aspects of *each* test will be briefly discussed in Appendix A.

Methodology

From August 2006 to April 2007 the research team used several methods to gather the information on each English language proficiency test. Initially, we searched state educational department and test developer/publisher websites and reviewed all documents pertinent to English language proficiency assessments for Title III purposes. Next, team members contacted Title III directors from all 50 states and the District of Columbia via e-mail. Title III directors were asked to share technical manuals, test administration manuals, alignment studies, website resources, and other relevant documents which would provide information on their state-adopted tests. Where necessary, team members also requested information from test developers/publishers and state assessment divisions. Follow-up phone calls were made to clarify test information, or to contact state representatives or publishers when prior attempts had not been successful. Informal email and phone conversations with test publishing companies/developers and state departments of education personnel served as additional sources of information in the data collection process.

The information provided in this chapter is as accurate and as current as possible at the time of publication. However, the following pages reflect only a “moment in time” snapshot of a dynamic process that is occurring nationwide as states and consortia continue their efforts to fully comply with English language proficiency assessment provisions within No Child Left Behind. In some cases, we had to rely upon unpublished and informal sources of data regarding assessment validity and reliability, standard setting, and item analysis. In many cases, states and test developers were still analyzing test data and/or technical manuals had not yet been published. For all of these reasons, changes to the technical data in this chapter are inevitable. Updates from the test developers and from state representatives can be incorporated in the web-based version of this report over time.[1]

Descriptions of Tests Currently Used by States for Title III Purposes

Summary information is provided for each assessment by test name, followed by grades covered, domains tested, publication date, and states using the assessment for Title III purposes. These summaries also include a description of the test purpose, score reporting information (i.e., proficiency levels), and a brief description of the test development. Where available, information about alignment activities and studies conducted during and after test development is included. Lastly, a discussion of the technical aspects of the test is included. Where available, information on item analysis, test reliability, validity, and freedom from bias is provided. The focus of the section on test technical properties is the types of psychometric tests conducted for each assessment; detailed results of each psychometric test are not provided. For more information on results of psychometric analysis for each assessment, the reader is referred to the test technical manual (as available).

As was indicated above, summary information for each individual assessment is provided in Appendix A. However, in this chapter, we present a summary that is characteristic of these assessments listed in Table 1 and discussed in Appendix A. Data from Table 1 will help the readers of this report gain a general idea of what ELP assessments are used by which states. Those who are interested in the details of some of these assessments may then read summary information on the assessment in Appendix A.

[1] See the UC Davis School of Education web site at http://education.ucdavis.edu/research/ELP_Assessment.html

RETURN to TABLE of CONTENTS

Texas English Language Proficiency Assessment System (TELPAS)

Grade Cluster(s): Reading Proficiency Test in English (RPTE): 3; 4–5; 6–8; and 9–12. Texas Observation Protocols (TOP): each individual grade from K to 12

Domains Tested: Reading Proficiency Test in English (RPTE): 3-12 reading. Texas Observation Protocols (TOP): K-2 reading; K-12 speaking, listening, and writing

Date(s) Published: Reading Proficiency Test in English (RPTE): 2000. Texas Observation Protocols (TOP): 2005

State(s) Using This Test for Title III Accountability (Implementation Date):
Texas (2005)

Test Purpose

The TELPAS system consists of two components: the Reading Proficiency Test in English (RPTE) and the Texas Observation Protocols (TOP). Since 2005, the TELPAS results have been used in the Annual Measurable Achievement Objective (AMAO) accountability measures required by NCLB. The TELPAS is used to measure students' annual progress in learning English in listening, speaking, reading, and writing. The TELPAS system is also used in combination with other measures to make instructional decisions for individual students. Beginning the 2007-2008 school year, only the TELPAS acronym will be used for RPTE and TOP.

Score Reporting

The four TELPAS proficiency ratings are: Beginning, Intermediate, Advanced, and Advanced High. Students are given a proficiency rating in reading, writing, listening and speaking. A comprehension rating is also given; the listening and reading ratings are each converted to a number from 1 (Beginning) to 4 (Advanced High). The average of the two numbers is the comprehension score. An overall composite level of proficiency, which combines the results of all four language domains, is also given. The language domain of reading is given most weight in the composite rating, followed by writing, listening and speaking have the least weight. The composite score ranges from 1 (ratings of Beginning in all language areas) to 4 (ratings of Advanced High in all language areas).

TOP is holistically scored; skills are not assessed in isolation. The TOP Proficiency Level Descriptors are the holistic scoring rubrics used by teachers to give one of four proficiency ratings in each of the four domains of reading, writing, listening and speaking for K-2 and listening, speaking and writing for grades 3-12.

Test Development

The RPTE was originally developed in response to Texas state regulations passed in 1995. Based on the recommendations of an advisory committee of assessment specialists and content experts, the Texas Education Agency (TEA) developed prototype test items in conjunction with Pearson Educational Measurement and Beck Evaluation and Testing Associates (BETA), the test contractors. The resulting items were field tested in the spring of 1999. In the fall of 1999, TEA conducted a field study to determine the test format and length. Following the spring 2000 test administration, raw score ranges for each proficiency level were established by TEA in conjunction with external assessment and content experts and practitioners based on second language acquisition theory and statistical analyses of student performance. Scaling of the assessment was conducted in fall 2000.

New items are written each year and reviewed by educators in the State of Texas. These items are then field tested in spring of each year. The TEA has undertaken the development of a second edition of RPTE beginning in the 2004–2005 school year. This second edition will add a second-grade assessment form and change the grade clusters to 2, 3, 4–5, 6–7, 8–9, and 10–12. This revised version will assess more of the type of reading required in the subject areas of science and mathematics. Field-testing of the second edition took place in spring 2006 and 2007, and the new edition will be implemented in spring 2008.

The Texas Education Agency (TEA), in conjunction with its testing contractor Pearson Educational Measurement, developed the TOP to assess the federally required domains and grade levels not tested on the RPTE. TOP was created by TEA along with test development contractors, bilingual/ESL consultants, and members of an English language learner focus group composed of teachers, bilingual/ESL directors, assessment directors, campus administrators, and university professors. TOP assesses students through observations in an authentic environment as students engage in regular classroom activities. In grades 2–12, the writing domain is assessed through a collection of classroom-based writing. The test was benchmarked in 2004 and fully implemented beginning in 2005.

RETURN to TABLE of CONTENTS

Standard Setting

The TEA and its testing contractors, technical experts and second language acquisition experts, an English language learner (ELL) assessment focus group of Texas educators and administrators from regional, district, and campus levels, and other Texas professional educators assisted in creating composite rating weighting formulas for the 2005 and 2006 TELPAS assessments to determine cut scores for each of the four proficiency levels within each domain and for the overall proficiency ratings. Additional information on scoring and standard setting is available in the technical report.

Alignment to State Standards

The RPTE was developed to align with the state's previous assessment program, the Texas Assessment of Academic Skills (TAAS). Beginning in spring 2004, RPTE was revalidated to be more closely aligned with the Texas Assessment of Knowledge and Skills (TAKS) reading selections and test questions. The TAKS, in turn, was developed to align with state content standards, providing a link between RPTE and Texas' content standards. The Texas Education Agency reports that the RPTE II, which will be fully implemented in 2008, will be aligned to the Texas content standards for reading and the English language proficiency standards, which emphasizes academic English acquisition.

Technical Properties of the Test

Item analysis. Each RPTE test question and reading selection is developed to align with proficiency level descriptors that are the foundation for test development and test construction. Before and after field testing, committees of educators review the reading selections and items to eliminate potential bias and ensure appropriateness in terms of content, age appropriateness, and proficiency level alignment. To determine the quality of each test item, the testing contractor produces statistical analyses for each using 3 types of differential item analyses: calibrated Rasch difficulty comparisons, Mantel-Haenszel Alpha and associated chi-square significance, and response distributions. Point biserial data are also evaluated yearly for each test item. Additionally, in order to ensure that the items perform consistent with the proficiency level descriptors and discriminate between students in the various proficiency level categories, the p-values of students in each proficiency level category are examined for each field-tested item. The educator review committees are provided with these data for each item that is field-tested in the annual field-test items review procedure. Using this information, item review committees review newly developed items for appropriateness of each item for use on future tests.

Test reliability. Internal consistency, the standard error of measurement (SEM) and the conditional SEM were calculated. Reliability estimates were also reported for items from the 2005–2006 test administration. Reliability is expressed in stratified alpha reliability for tests/objectives involving short-answer and/or essay questions; KR-20 reliability was computed for all other question types. These reliabilities are provided by grade and by grade/gender. Reliability coefficients are reported for grades 3, 4-5, 6-8, and 9-12.

A large-scale study of rating validity and reliability of the TOP was conducted by TEA in spring of 2006. An audit of more than 13,000 scored writing samples collected from teachers who were trained TOP raters was conducted to evaluate how effectively raters applied the rubrics. Individuals trained as TOP raters at the state level rescored the student writing collections. Overall the state and teacher ratings agreed perfectly 77% of the time. The study also required the raters of the students selected for the audit to complete a questionnaire concerning the adequacy of the training and scoring processes for each language domain. Of the more than 6,000 raters audited, following are the percents of raters indicating that the training provided them with sufficient information to judge the English language proficiency levels of their students in each language domain: listening 96%, speaking 96%, writing 97%, and reading (grade 2 only) 94%. Detailed information on this study is available in the technical report.

Test validity. Two studies examined the relationship between RPTE and TAKS performance levels. The first study, which took place after the spring 2004 test administration examined the following issues: 1) the percent of qualifying recent immigrants who met the AYP incremental progress performance standard in spring 2004, and 2) the reading performance of LEP students evaluated under the incremental progress model compared to that of LEP students evaluated with TAKS and 3) the instructional rationale for incremental RPTE progress model. These statistical alignment analyses indicated that the percentages of students who met the RPTE incremental progress standard and TAKS standard were very similar. A second study un-

RETURN to TABLE of CONTENTS

dertaken after the spring 2005 test administration established a connection between RPTE scores and the TAKS performance categories of Met Standard (passing level) and Commended Performance (highest performance level). In addition, content validation studies are conducted yearly by panels of Texas teachers, test development specialists and TEA staff members. Specific information on test validity is given in the technical digest.

Freedom from bias. Please see technical manual for details of how freedom from bias issues were addressed for each test in the TELPAS system.

The summary given above was an example of the ELP tests used by one state. Similar data are presented for other states in Appendix A. Once again, it must be indicated that information on currently-used Title III ELP assessments is time sensitive and is subject for frequent revisions. We plan to revise this chapter and the related information in Appendix A as soon as we receive them from states.

Technical Reports

Technical Digest 2004–2005. *Student assessment division*. Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Technical Digest Texas English Language Proficiency Assessment System (TELPAS) 2004–2005. Appendix 7: *Development of the TELPAS composite ratings and composite scores*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/ListofAppendices/TechDigest-A7.pdf

Texas Assessment. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Education Agency (2002). *Student Assessment Division: Technical Digest 2001–2002*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig02/index.html

Texas Education Agency (2003). *Student Assessment Division: Technical Digest 2002–2003*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig/index.html

Texas Education Agency (2004). *Student Assessment Division: Technical Digest 2003–2004*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig04/index.html

Texas Education Agency (2005). *Student Assessment Division: Technical Digest 2004–2005*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Texas Assessment (2006a). *Student assessment division: Technical digest 2005–2006*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Assessment (2006b). Student Assessment Division: Technical Digest 2005–2006. *Appendix 6*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Assessment (2006c). Student Assessment Division: Technical Digest 2005–2006. *Chapter 15: Validity*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/Chapters/Chapter15_Validity.pdf

Texas Assessment (2006d). Student Assessment Division: Technical Digest 2005–2006. *Chapter 17: Reliability*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/Chapters/Chapter14_Reliability.pdf

Texas Assessment (2006e). Student Assessment Division: Technical Digest 2005–2006. *Appendix 10*. Retrieved September 3, 2007 from: http://k12testing.tx.ncspearson.com/TechDigest/ListofAppendices/TechDigest-A10.pdf

Technical Digest 2004–2005. *Student assessment division*. Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Technical Digest 2004–2005. Appendix 7: *Development of the TELPAS composite ratings and composite scores*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/ListofAppendices/TechDigest-A7.pdf

Texas Assessment. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

TABLE 1. *Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007*

State	First Implemented	Name of Test	Test Developer
Alabama	Spring 2005	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Alaska	Spring 2006	IPT® Title III Testing System (IPT) [LAS (Forte, 2007)]	Ballard & Tighe
Arizona	Fall 2006	Arizona English Language Learner Assessment (AZELLA)	Arizona Department of Education; Harcourt Assessment Inc.
Arkansas	Spring 2007	English Language Development Assessment (ELDA) English Language Development (ELDA) K–2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
California	Fall 2001	California English Language Development Test (CELDT)	California Department of Education; CTB/McGraw Hill
Colorado	Spring 2006	Colorado English Language Assessment (CELA)	CTB/McGraw Hill
Connecticut	Winter & Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Delaware	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Florida	Fall 2006	Comprehensive English Language Learning Assessment (CELLA)	Accountability Works; Educational Testing Service (ETS); and a consortium of 5 states
Georgia	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Hawaii	Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Idaho	Spring 2006	Idaho English Language Assessment (IELA)	Questar Assessment, Inc. (formerly Touchstone Applied Science Associates)
Illinois	Spring 2006	Assessing Comprehension and State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Indiana	Winter and Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Iowa	Spring 2006	English Language Development Assessment (ELDA) English Language Development (ELDA) K–2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
Kansas	Spring 2006	Kansas English Language Proficiency Assessment (KELPA)	The Center for Testing and Evaluation (CETE); Kansas State Department of Education; University of Kansas
Kentucky	Spring 2007	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)

Louisiana	1. Spring 2005 (grades 3-12)
	2. Spring 2006 (grades K-2 added)
		1. English Language Development Assessment (ELDA)
		2. English Language Development (ELDA) K–2 Assessment
			American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE);
			Council of Chief State School Officers (CCSSO); Measurement Inc.(MI);
			State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)

Maine	Spring 2005	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Maryland	Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Massachusetts	1. Spring 2004 2. Spring 2004 3. Spring 2007	1. Massachusetts English Proficiency Assessment-Reading & Writing (MEPA-R/W) 2. Massachusetts English Language Assessment-Oral (MELA-O) 3. IPT® 2004:IPT Early Literacy Test reading and writing (K-2 reading and writing only)	1. Massachusetts Department of Education; Measured Progress 2. Educational Assistance Center (EAC) East; Massachusetts Assessment Advisory Group (MAAG); Massachusetts Department of Education 3. Ballard & Tighe
Michigan	2006	Michigan English Language Proficiency Assessment (MI-ELPA)	Harcourt Assessment Inc.; Michigan Department of Education
Minnesota	1. Fall 2001 2. 2002 – 2003 academic year	1. Test of Emerging Academic English (TEAE) 2. MN SOLOM	1.MetriTech, Inc.; Minnesota Department of Education 2. Bilingual Education Office of the California Department of Education; San Jose Area Bilingual Consortium
Mississippi	2003 - 2004 academic year	The Stanford English Language Proficiency Test (SELP, Stanford ELP)	Harcourt Assessment Inc.
Missouri	Winter 2002	Maculaitis Assessment of Competencies II (MAC II)	Questar Assessment, Inc. (formerly Touchstone Applied Science Associates)
Montana	Winter 2006	MontCAS English Language Proficiency Assessment (MONTCAS ELP)	Measured Progress; Mountain West Assessment Consortium (MWAC); Questar Assessment, Inc. (formerly Touchstone Applied Science Associates) has taken over production of test

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
Nebraska	Spring 2005 (grades 3-12) Spring 2006 (grades K-2 added)	English Language Development Assessment (ELDA) English Language Development (ELDA) K–2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Nevada	2005 - 2006 academic year	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
New Hampshire	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
New Jersey	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
New Mexico	Spring 2006	New Mexico English Language Proficiency Assessment (NMELPA)	Harcourt Assessment Inc.; New Mexico Department of Education
New York	Spring 2005	New York State English as a Second Language Achievement Test (NYSESLAT)	Educational Testing Service (ETS); Harcourt Assessment Inc.; New York State Education Department
North Carolina	2005	IPT® Title III Testing System (IPT)	Ballard & Tighe
North Dakota	Spring 2007	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Ohio	1. Spring 2006 (grades 3-12) 2. Spring 2006 (K-2 only)	1. Ohio Test of Language Acquisition (OTELA) 2. English Language Development (ELDA) K-2 Assessment	1. American Institutes for Research (AIR); Ohio Department of Education 2. American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Oklahoma	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Oregon	Spring 2006	Oregon English Language Proficiency Assessment (ELPA) [SELP (Forte, 2007)]	Language Learning Solutions (LLS)
Pennsylvania	Spring 2007	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
Rhode Island	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
South Carolina	Spring 2005 (Grades 3-12) Spring 2006 (Grades K-2 added)	English Language Development Assessment (ELDA) English Language Development (ELDA) K-2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
South Dakota	Spring 2006	Dakota English Language Proficiency Assessment (Dakota ELP)	Harcourt Assessment Inc.; South Dakota Department of Education
Tennessee	1. Spring 2005 2. 2007 3. 2007	1. Comprehensive English Language Learning Assessment (CELLA) 2. English Language Development Assessment (ELDA) 3. English Language Development (ELDA) K-2 Assessment	1. Accountability Works; Educational Testing Service (ETS); and a consortium of 5 states 2./3. American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Texas	 A. 2000 B. 2005	1. Texas English Language Proficiency Assessment System (TELPAS) A. Reading Proficiency Test in English (RPTE) B. Texas Observation Protocols (TOP)	Beck Evaluation and Testing Associates (BETA); Pearson Educational Measurement; Texas Education Agency (TEA)
Utah	Fall 2006	Utah Academic Language Proficiency Assessment (UALPA)	Measured Progress; Mountain West Assessment Consortium (MWAC)
Vermont	Spring 2005	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Virginia	Spring 2006	Virginia Stanford English Language Proficiency Test	Harcourt Assessment, Inc.
Washington	2006	Washington Language Proficiency Test II (WLPT-II)	Harcourt Assessment Inc.; Washington Department of Education
Washington D.C.	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
West Virginia	1. 2005	1. West Virginia Test for English Language Learning (WESTELL)	1. N/A
	2. Spring 2005 (grades 3-12)	2. English Language Development Assessment (ELDA)	2./3. American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
	3. Spring 2006 (grades K-2 added)	3. English Language Development (ELDA) K-2 Assessment [ELDA only (Forte, 2007)]	
Wisconsin	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Wyoming	Spring 2006	Wyoming English Language Learner Assessment (WELLA)	Harcourt Assessment Inc.; Wyoming Department of Education

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

Chapter 8

Establishing and Utilizing an NCLB Title III Accountability System: California's Approach and Findings to Date

Robert Linquanti and Cathy George[1]

The federal No Child Left Behind (NCLB) Act of 2001 inaugurated major changes in the expectations placed on state and local education agencies regarding assessment of and accountability for the performance of Limited English Proficient (LEP) students—also known as English Language Learners (ELLs) (NCLB, 2002). Specifically, NCLB Title III requires states to: 1) establish English language proficiency (ELP) standards aligned to state academic content standards yet suitable for ELLs learning English as a second language; 2) annually assess the English-language proficiency of each ELL using a valid and reliable assessment of English-language proficiency aligned to ELP standards; 3) define annual measurable achievement objectives (AMAOs) to measure and report on progress toward and attainment of English proficiency and academic achievement standards; and 4) hold local education agencies (LEAs) accountable for meeting increasing AMAO targets over time (NCLB, op. cit.).

These new mandates have generated significant challenges for states with respect to standards and test development, test validity, and accountability policy development and implementation (GAO, 2006; Zehr, 2006; Abedi, 2004; Crawford, 2002). Previous chapters in this volume explore and discuss challenges and strategies in defining ELP standards and developing and implementing valid and reliable standards-based ELP assessments. Beyond issues of ELP test development and implementation, states also face complex technical and policy issues in using ELP assessment data to define AMAOs and establish accountability systems under Title III (GAO, op.cit.). Some of these issues include determining reasonable annual growth expectations in English; operationally defining the English proficient level; and setting baselines and annual growth targets for local education agencies (George, Linquanti & Mayer, 2004; Gottlieb & Boals, 2005; Linquanti, 2004). Additionally, Title III accountability systems need to be designed carefully so that they are understood and supported by policymakers, local educators, students, and the community. These technical and policy issues are the focus of this chapter.

Specifically, this chapter presents and discusses methods that a team of California Department of Education

[1] The ideas and opinions presented in this chapter are those of the authors and do not necessarily reflect those of their respective agencies.

RETURN to TABLE of CONTENTS

(CDE) staff and outside technical consultants[2] used to develop and implement Title III accountability in the nation's largest ELL-enrolling state. Because California already had its English language proficiency test in place for two years when it developed its Title III accountability system in 2003, the AMAO development team was able to model potential outcomes of different policy options using matched-score results for over 862,000 ELL students. This chapter reviews empirical methods used to define progress expectations for learning English under AMAO 1 and for attaining English language proficiency under AMAO 2—the two ELP-related AMAOs; and to establish AMAO starting points, ending points, and annual growth target structures. The chapter then reports on three years of subsequent ELP AMAO data analyses using four years of California English Language Development Test (CELDT) results on over one million ELLs annually to examine actual-to-expected progress. Finally, the chapter examines accountability outcomes for LEAs to date using results on all three AMAOs—including AMAO 3, the AMAO that measures annual yearly progress (AYP) of ELLs on academic achievement assessments.

A Brief Note on the California English Language Development Test (CELDT)

California had a unique advantage when it developed its Title III accountability system in 2003. It already had a standards-based English language proficiency assessment in place. The CELDT, based on ELP standards drafted by a statewide panel of practitioners and experts in 1999–2000, became operational in California in 2001. The CELDT has test forms for each of four grade spans: K–2, 3–5, 6–8, and 9–12. An overall composite score was derived from three separate domain scores[3] which were weighted as follows: listening & speaking (50%); reading (25%); and writing (25%). CELDT has five overall proficiency levels: *beginning, early intermediate, intermediate, early advanced,* and *advanced*. The test has changed over time, and will continue to do so as California addresses all the requirements for English language proficiency assessment under Title III.[4] It is beyond the scope and not the purpose of this chapter to discuss the development and technical qualities of the CELDT. There is an ongoing program of research and development for the test, and test blueprints and technical reports are publicly available.[5] Finally, as part of the CELDT administration, the local education agency (LEA) provides the ELL student's CELDT scores from the previous year and reports the year the student was first enrolled in school in the United States. As will be seen, these variables are necessary for the calculation of AMAO results.

Development of AMAO 1 (Progress toward English-Language Proficiency)

The first AMAO relates to making progress in learning English. Title III required that states determine annual increases in the number or percentage of students who make progress in learning English as measured by the state's English language proficiency test. There were several key decisions that needed to be made in order to establish AMAO 1:

- Determine the scoring metric to be used to measure growth
- Determine the annual growth target
- Set the starting point for AMAO 1 targets (2003–04)
- Set the ending point for AMAO 1 targets (2013–14)
- Determine the rate of annual growth from 2004 to 2014

Because AMAO 1 considers annual progress in English language development, it requires that each student included in the cohort for the analysis be tested at least two points in time so that a growth score can be calculated. In California the CELDT is used for identification of ELL status, and is administered within 30 days of initial enrollment in a

[2] The authors wish to acknowledge their fellow core team members: Jan Mayer, State Title III Director in 2003; Gloria Cardenas, Consultant at CDE, and Hiro Hikawa, Statistician at American Institutes for Research.

[3] Although California's ELP standards encompass all four domains, reading and writing are tested in grades 2–12, while listening & speaking are assessed in grades K–12. Beginning with the 2006–07 form, the CELDT weights are listening (25%); speaking (25%); reading (25%) and writing (25%).

[4] For example, CDE has recently developed a comprehension score and separate listening and speaking scores, and is currently seeking legislative authority to assess reading and writing in kindergarten and grade one.

[5] See the CELDT web site at http://www.cde.ca.gov/ta/tg/el/resources.asp.

RETURN to TABLE of CONTENTS

California school and annually thereafter until a student is re-designated as fluent English proficient (R-FEP).

Determining the Scoring Metric

One of the first decisions that had to be made in setting AMAOs 1 and 2 was the choice of the scoring metric to be used. The CELDT yielded three score types: Raw scores, scaled scores, and proficiency scores.

Raw scores were eliminated from consideration because changes in raw scores would have inconsistent meanings and would be exceedingly hard to interpret. The decision was therefore focused on whether to use proficiency scores or scaled scores. The use of scaled scores was examined, as this might allow for equal intervals of growth to be measured from any point in the continuum of English language proficiency. However, it was not possible to use scaled scores on the CELDT because, at the time of the development of the English language proficiency AMAOs, the scores from the different grade span tests were not vertically equated and could not be compared.[6] Therefore, proficiency scores were found to be the most viable metric for growth. At the time the Title III accountability system was developed, CELDT yielded the following proficiency scores: a combined listening/speaking score, reading score, writing score, and a combined overall proficiency score. The reliability of the subskill (domain) scores was not considered sufficiently strong to set progress expectations at the subskill level. Therefore, the *overall* proficiency score was used as the scoring metric for AMAO 1.

Determining the Annual Growth Target

Using overall proficiency scores as the metric of growth limited the range of options in specifying how much growth should be expected in one year. An obvious target was to expect students to gain one proficiency level per year. Since NCLB Title III does not require 100% of ELLs to make progress learning English in any given year, it was feasible to consider such an annual growth target and specify the minimum percentage of ELLs required to meet it. Empirical data on the matched score sample of over 862,000 ELL students from the 2000 and 2001 administrations of the CELDT demonstrated that 50% of students gained one or more proficiency levels from 2001 to 2002, 40% remained at the same proficiency level, and 10% dropped one or more proficiency levels. Patterns in proficiency level growth varied by proficiency level and by grade. The greatest gains were made at the *beginning* and *early intermediate* levels where 70% and 62% of students, respectively, gained one or more proficiency levels. Growth at the *intermediate* levels and above was much more difficult. At the *intermediate* level, 44% of the students gained one level. Only 26% of students at *early advanced* level gained a level that year. The students at the *advanced* level had reached the highest proficiency level and could only remain at that level—which 42% did—or decrease in proficiency—which 58% did. Similar patterns have been seen in other ELP assessments, both old and new (e.g., de Avila, 1997; WIDA Consortium, 2006). Because AMAO 1 needed to include a growth target for all students, even the students at the highest proficiency levels of the CELDT, it was recommended that if students had reached the level on CELDT considered sufficient for reclassification, they should maintain that level until reclassified.

In California, reaching the *English proficient* level on CELDT is only one criterion that ELLs must satisfy to be considered for reclassification. Other criteria include academic achievement on the California Standards Test, teacher judgment, and parent input. A student is defined as *English proficien*t on the CELDT if both of the following are met:

- Overall proficiency level score is *early advanced* or *advanced* and
- Each skill area proficiency score is at the *intermediate* level or above.

The growth target was set so that students are expected to gain one overall proficiency level each year until they reach the e*arly advanced* or *advanced* level overall. Those at the *early advanced* or *advanced* level who are not yet English proficient are expected to achieve the *English proficient* level on CELDT (i.e., bring all of their subskills to the *intermediate* level or above). Those previously at the *English proficient* level on CELDT, but not meeting other reclassification criteria, take the CELDT again and are expected to maintain the *English proficient* level. Figure 1 summarizes the AMAO 1 annual growth target rules. Using these AMAO 1 growth target rules, outcomes for LEAs were modeled using the 2000–2001 empirical data set to investigate the possibility of disparate impact on elementary versus secondary versus

[6] A common scale for the CELDT has recently been developed, and in the future it may be possible to use scaled scores as a metric of growth on CELDT for Title III accountability.

RETURN to TABLE of CONTENTS

Figure 1. AMAO 1 annual growth targets (using CELDT proficiency levels)

Previous Year CELDT Overall Proficiency Score	Annual Growth Target
• *Beginning* → • *Early intermediate* → • *Intermediate* →	• *Early intermediate* Overall • *Intermediate* Overall • *Early advanced* Overall
• *Early advanced/advanced*, but not at *English proficient* level (i.e., one or more skill areas below *Intermediate*) →	• Achieve *English proficient* level. (Overall *early advanced/advanced* and all skill areas at *intermediate* level or above)
• *Early advanced* or *advanced* and at the *English proficient* level →	• Maintain *English proficient* level

unified school districts. No notable disparate impact was found.

Finally, the AMAO development team considered further elaborating annual growth targets by grade and grade span. Using the matched-score sample, changes in proficiency level were examined across grades (e.g., K–1, 1–2, etc.) as well as within and across grade spans (K–2, 3–5, etc.). As has been observed in other ELP assessments based on standards defined by grade spans (see *Overview of Existing English Language Proficiency Tests* chapter in this report), a greater percentage of students tended to maintain or decrease in overall proficiency level when they crossed a grade span. (This very likely reflects the increased difficulty in test items that are based on standards for higher grade levels.) Furthermore, the fact that reading and writing assessments are introduced on the CELDT in grade 2 further complicated performance patterns within this grade span. As a result, the AMAO team chose not to further elaborate annual growth targets based on a student's grade or grade span.

Setting the Starting Point for AMAO 1 Targets

Once the annual growth target was established for AMAO 1, it was necessary to determine what percentage of students within each LEA would be required to meet the growth target. The percentage of students meeting the growth target was analyzed for all LEAs having 25 or more ELL students with the necessary two years of CELDT data. Title I AYP offered a procedure for determining the starting point that was modeled at the LEA level for Title III. (Recall that in Title III, LEAs are used instead of schools because LEAs are held accountable, not schools.) In the Title I method, schools are ranked from low to high according to the percentage of students achieving the growth target and the performance of the school at the 20th percentile of the state's distribution is used as the starting point. Applying this method, 20% of Title III LEAs would be below the target selected, and 80% would meet or exceed the target. This was judged to be a reasonably ambitious starting point as it signaled that the lowest 20% of LEAs would need to immediately improve their performance in helping ELLs to progress annually. This method of determining the starting point resulted in a starting target of 51% of students within each LEA being expected to meet their annual growth target.

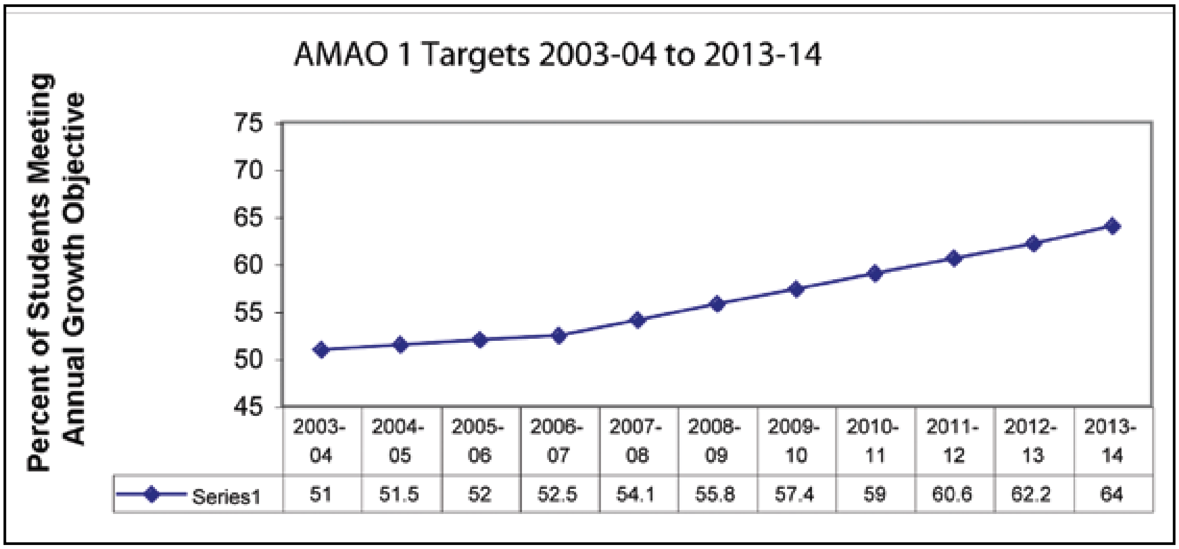
Setting the Ending Point for AMAO 1 Targets

Title III requires annual increases in the percentage of students making progress in learning English. However, under Title III, states were not required to set the ending target in 2013–14 at 100% as was the case with Title I AYP. This allowed for the establishment of a more reasonable ending target that was based on empirical data. In California, three scenarios were considered: an ending target at the 60th percentile of the 2001–2002 LEA distribution; at the 75th percentile of that distribution; and at the 90th percentile of that distribution. After consideration of these alternatives, the AMAO team recommended—and state policymakers decided—that by 2014 all LEAs should reach the point that the top 25% of LEAs (i.e., 75th percentile of the distribution) had attained in 2001–02 (i.e., 64% of ELLs making progress). This seemed an attainable yet rigorous target for LEAs to meet.

Once the starting and ending points had been established, the intervals of growth needed to be determined. Smaller increments of growth were established for the first three years to allow districts to become accustomed to the accountability system and allow time for improved instructional practices and systems. After the first three years, growth was set at equal intervals for the next seven years (see Figure 2).

RETURN to TABLE of CONTENTS

Figure 2. AMAO 1 targets for Title III–funded LEAs in California



Development of AMAO 2: Attaining English-Language Proficiency

The second AMAO relates to ELLs attaining English-language proficiency. Title III requires that states determine annual increases in the number or percentage of students who attain English language proficiency as measured by the state's English language proficiency test. As with AMAO 1, several key decisions needed to be made in order to establish AMAO 2:

- Define the *English proficient* level
- Determine the cohort of ELLs for analysis
- Set the starting point for AMAO 2 targets (2003–04)
- Set the ending point for AMAO 2 targets (2013–14)
- Determine the rate of annual growth from 2004 to 2014

Defining the English Proficient Level

The AMAO development team operationalized and tested various definitions of the *English proficient* level on CELDT using empirical data. The definitions tested included: a) *early advanced* or higher proficiency level overall with all subskills at *intermediate* or higher; b) *early advanced* or higher proficiency level overall with all subskills at *early advanced* or higher; and c) *early advanced* or higher proficiency level overall with **no minimum level** on subskills. While there was a relatively small difference in outcomes between the first two definitions (with definition b. slightly more difficult to attain), both a. and b. were substantially more challenging to attain than definition c. The AMAO team therefore recommended keeping the same definition of English language proficiency using CELDT already adopted by the State Board of Education (definition a.), since: it was judged to be sufficiently rigorous; it would complement the definition used for AMAO 1 and reclassification; and it would allow the Title III accountability system be consistent with existing state guidelines, which California educators understood and had largely adopted, thus increasing the likelihood of the accountability system's acceptance statewide.

Determining the Cohort of ELLs for Analysis

NCLB Title III requires that AMAOs be developed in a manner that reflects the amount of time an individual child has been enrolled in a language instruction educational program. This AMAO therefore entails a cohort analysis. One key issue to address is *which ELL students can reasonably be expected to reach English language proficiency at a given point in time*—which effectively determines the denominator for the AMAO 2 calculation. This is optimally determined using longitudinal data, in order to propose targets for students based on: 1) their English language

RETURN to TABLE of CONTENTS

proficiency levels when they enter U.S. schools, and 2) their corresponding attainment of the *English proficient* level over time. However, given significant data limitations, the AMAO development team examined students with two CELDT data points and information regarding their length of time in U.S. schools.

Various cohort definitions were considered and modeled with the existing empirical data. Given the annual growth target already defined under AMAO 1 (see above), it appeared defensible to set a four-year criterion as one key factor for students' inclusion in the AMAO 2 cohort. Existing empirical studies of the time needed to attain language proficiency, which estimate three to five years for oral fluency (De Avila, 1997) and four to seven years for overall English-language proficiency (Hakuta et al., 2000), also lent support to establishing a four-year criterion for cohort inclusion. (Recall that Title III does *not* stipulate—as in Title I—that 100% of students included in AMAO 1 or AMAO 2 calculations need to make progress or attain the *English proficient* level in any given year. This allowed for ambitious yet realistic targets to be set based on available empirical data.) In addition to years in U.S. schools, another key factor considered in determining which students to include for analysis was students' prior CELDT level since this may also indicate which students can reasonably be expected to reach English language proficiency from one year to the next. This includes students enrolled for a shorter period of time but who may have entered the system with a higher level of initial language proficiency and for whom attaining the *English proficient* level is a reasonable expectation.

Our empirical data analyses showed that about 44% of ELLs in U.S. schools for four or more years attained the *English proficient* level in 2002–03. Moreover, 41% of ELLs with *intermediate* proficiency in 2001–02 attained the *English proficient* level in 2002–03. Conversely, about 2% of ELL students at *early advanced* or *advanced* proficiency in 2001–02 still did not meet the *English proficient* level in 2002–03 due to one or more subskill scores remaining below *intermediate*. Finally, 14% of ELLs below *intermediate* in 2001–02 reached the *English proficien*t level in 2002–03. Each of these categories contributed to the cohort definition ultimately adopted.

It was also decided to consider as attaining the *English proficient* level only those students who "crossed the finish line" from not *English proficient* in the previous year to *English proficient* in the current year. Unlike AMAO 1, in AMAO 2 LEAs were not given credit for students remaining at the *English proficient* level. In this way, the accountability system would not create a perverse incentive for LEAs to keep high-performing CELDT test-takers classified as ELLs in order to boost the percentage scoring at the *English proficient* level year after year.[7]

Several options for determining which students to include in the AMAO 2 cohort, along with the advantages and disadvantages of each option, were provided to the California State Board of Education. The final ELL cohort defined for AMAO 2 combines the following four groups of students:

- ELLs who were at the *intermediate* level in the prior year
- ELLs who were at *early advanced* and *advanced* but not *English proficient* (based on subskills) in the prior year
- ELLs who were at *beginning* or *early intermediate* in the prior year and who first enrolled in U.S. schools four or more years ago[8]
- ELLs who were at *beginning* or *early intermediate* in the prior year and in U.S. schools *less than* four years who reach the *English proficient* level in the current year.

ELL students meeting any of these criteria are included in the cohort for AMAO 2 calculations.

Setting the Starting Point for AMAO 2 Targets

Once the cohort for analysis was established for AMAO 2, it was necessary to determine what percentage of students in the cohort—as a starting point within each LEA—would be required to meet the *English proficient* target. The percentage of ELL students in the AMAO 2 cohort

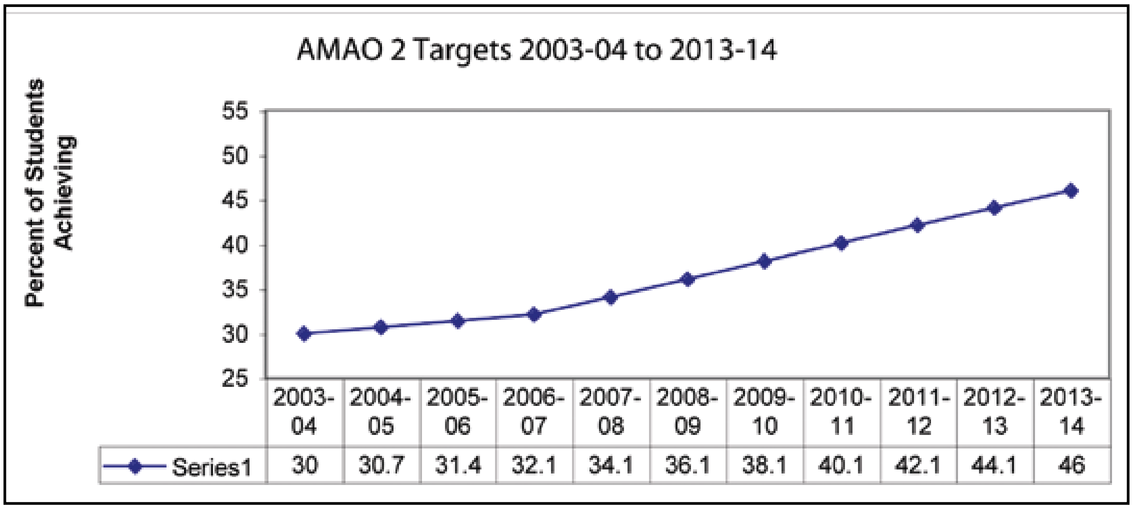
[7] Given California's multiple criteria—linguistic and academic—for English learner reclassification, the state's Title III accountability system considers as progressing under AMAO 1 those CELDT test takers maintaining the *English proficient* level while they pursue meeting other reclassification criteria.

[8] Since CELDT is administered in summer/fall of each school year, the four-or-more year enrollment requirement is operationalized as more than four years enrolled in U.S. schools.

RETURN to TABLE of CONTENTS

Chapter 8

Figure 3. AMAO 2 targets for Title III–funded LEAs in California



	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14
Series1	30	30.7	31.4	32.1	34.1	36.1	38.1	40.1	42.1	44.1	46

described above that attained the CELDT *English proficient* level was analyzed for all LEAs having 25 or more ELL students with the necessary two years of CELDT data and time-in-U.S.-schools information. As with AMAO 1, the Title I AYP procedure for determining the starting point was utilized for Title III: LEAs were rank-ordered based on the percentage of students attaining the *English proficient* level, and the AMAO 2 result of the LEA at the 20th percentile of the state's distribution was chosen as the starting point. By definition, 20% of Title III LEAs were below that target and 80% would meet the target. This was judged to be an ambitious yet defensible starting point. Using this method, 30% of students within each LEA's AMAO 2 cohort were expected to attain the *English proficient* level in 2003–04.

Setting the Ending Point for AMAO 2 Targets

Title III requires annual increases in the percentage of students making progress in attaining English language proficiency. However, as noted above, Title III does not require states to set the ending target in 2013–14 at 100% as was the case with Title I AYP. This allowed ambitious yet reasonable ending targets to be established based on empirical data. In California, three scenarios were considered: an ending target at the 60th percentile of the 2001–2002 rank-ordered LEA AMAO 2 performance distribution, at the 75th percentile of that distribution, and at the 90th percentile of the distribution. After considering these alternatives, California's State Board of Education decided that in 10 years all LEAs should reach the point that the top 25% of LEAs had attained at the starting point (i.e., 46% of the AMAO 2 cohort attaining the *English proficient* level). This was judged to be an attainable yet rigorous target for all of the state's LEAs to strive for.

Once the starting point and ending point had been established, growth intervals were set. Applying the method used in AMAO 1 targets, smaller growth increments for AMAO 2 were established for the first three years to allow districts to become accustomed to the accountability system and to implement instructional improvements. After the first three years, growth was set at equal intervals for the next seven years (see Figure 3).

Inclusion of AMAO 3: Academic Achievement of the ELL Subgroup per Title I Adequate Yearly Progress (AYP)

Under Title III, states are also required to measure the adequate yearly progress of ELL students attaining academic proficiency in core subject matter. The law stipulates that this measure of adequate yearly progress shall come from Title I for the ELL subgroup at the LEA level. AMAO 3—often referred to as the AYP AMAO—therefore holds LEAs accountable for the ELL subgroup meeting the same academic achievement targets that are required of all schools, districts and subgroups under NCLB Title I. There are four components to AMAO 3:

- participation rate in the English language arts (ELA) assessment

RETURN to TABLE of CONTENTS

- participation rate in the mathematics assessment
- percentage of subgroup that is proficient or above in ELA
- percentage of subgroup that is proficient or above in mathematics

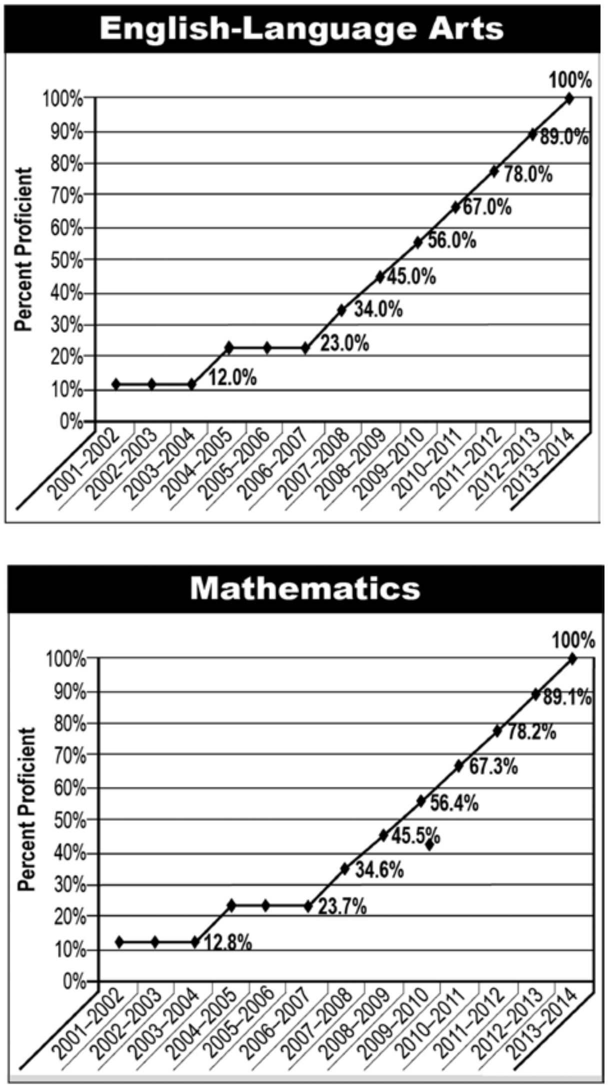
Federal participation-rate targets in ELA and mathematics specify that 95% of the students overall and in each significant subgroup must be tested.[9]

The academic targets for the percentage of students that are required to be proficient or above in ELA and mathematics are determined by each state and are specified in the state's accountability workbook for NCLB. Under NCLB the percentage-proficient targets for all schools, districts and significant subgroups must reach 100% in 2014. Unlike AMAOs 1 and 2, which are based on the *progress over time* that ELLs are making in attaining English proficiency, AMAO 3 assesses the *status at one point in time* of the ELL subgroup (i.e., no matched scores are used). The academic target is particularly difficult for the English learner subgroup because, by definition, these students are not proficient in English and the tests in California that are used to determine academic proficiency are administered in English. Federal regulations allow for the inclusion of some R-FEP (by definition higher-performing, former ELL) students in the ELL subgroup calculation, which somewhat mitigates the "skimming bias" caused by higher performing ELLs leaving the ELL subgroup category. Nevertheless, the ELL subgroup is being continually refreshed with new, lower-performing English learners entering schools in the United States, which negatively biases this subgroup's AYP results (see Abedi, 2004; and Linquanti, 2001).

In order for an LEA to meet AMAO 3 they must meet all four of the components listed above. The participation-rate targets are typically met by most LEAs; the academic targets in English language arts and mathematics are more difficult to meet and will become increasingly difficult as these targets approach 100% proficient. Figure 4 displays California's Title I AYP targets for ELA and math.

In order to be identified as a Program Improvement (PI) district under Title I, the same content area must be missed for 2 consecutive years, and in some states additional grade-span analyses are also used to limit the identification of districts as PI. However, these more restrictive conditions do not apply under AMAO 3. Hence more California districts fail to meet AMAO 3 under Title III than are identified as a PI districts under Title I AYP due to ELL subgroup performance, using the same target structure. As is seen in the next section, this third AMAO has had a disproportionate effect on Title III accountability outcomes.

Figure 4: Title III AMAO 3/Title I AYP Targets for CA Unified School Districts



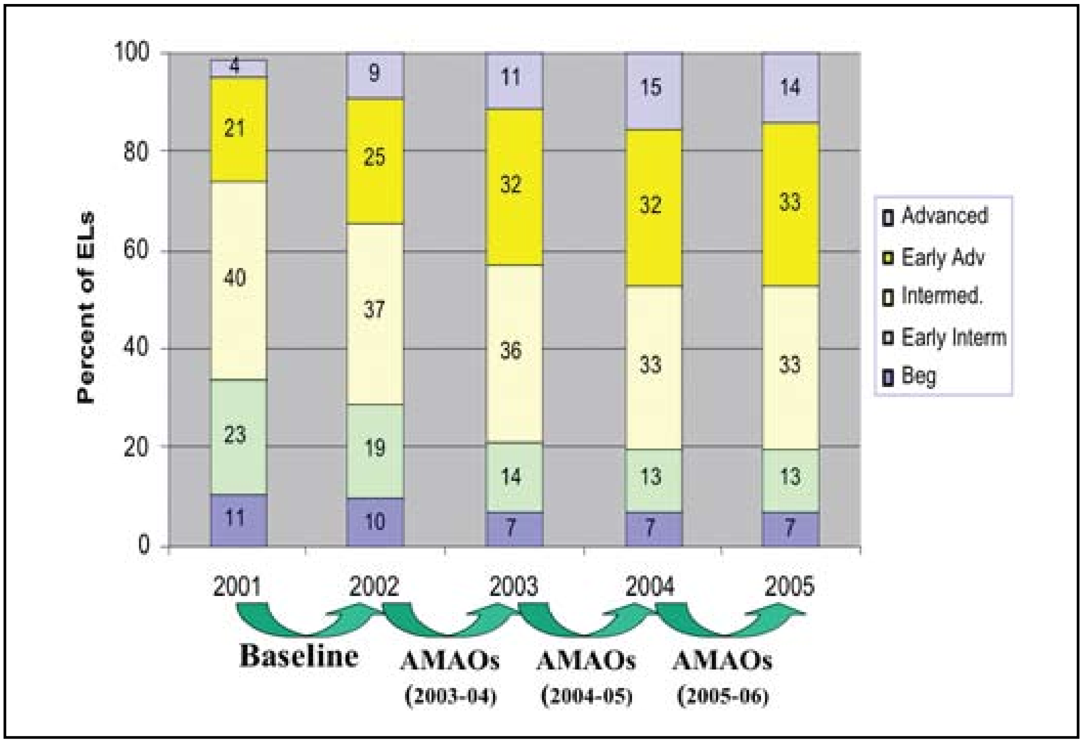
English-Language Arts

Mathematics

[9] While NCLB Title III does not require ELLs enrolled in a U.S. school for fewer than 12 months to be tested in English language arts, California law requires all ELLs to be tested regardless of length of time enrolled. However, scores of ELLs enrolled fewer than 12 months are not used to calculate Title I participation rates and AYP.

RETURN to TABLE of CONTENTS

Figure 5. CELDT proficiency level distributions and AMAO source data, by year.



How Has California's Title III Accountability System Performed: Outcomes to Date

The preceding sections discussed in detail how California's Title III accountability system was established. This accountability system has been in place since 2003, the longest of any state whose AMAO targets were based on empirical data on the population of interest. This section shares and discusses results to date from this accountability system.

Figure 5 shows the cross-sectional percentage distribution of performance by overall English proficiency level of English learner students on the annual CELDT test from 2001 to 2005. As discussed previously, a matched score sample of over 862,000 ELLs from the first two years of CELDT administration (2001 and 2002) was used as a baseline to establish metrics, growth targets, and beginning and ending points. Year-to-year matched score comparisons over the next three test administrations (2003 to 2005) were used to generate three years of AMAO results for Title III LEAs, as indicated in Figure 5 by the linked arrows. Since CELDT is administered in the summer/fall (July–October) of each school year, results are compared fall-to-fall, and AMAO 1 and 2 preliminary data are generated and provided to LEAs in the spring. Final judgment about whether LEAs satisfy all AMAOs is provided once AMAO 3 results are available from Title I AYP calculations, which are performed in summer after the spring administration of California's academic achievement tests. As Figure 5 shows, the proportion of ELL students reaching *early advanced/advanced* levels has increased from the early years, and held fairly steady at 47% during the past two years.[10] About one third of annual CELDT takers score at the overall *intermediate* level, while about one fifth score at *beginning* or *early intermediate* levels.

[10] The percentage of *early advanced/advanced* students also meeting the English-language proficient criterion—all subskills being *intermediate* or greater—is consistently about two to three percentage points lower. For example, for both 2004 and 2005, 47% are *early advanced/advanced,* while 44% are *English proficient* on CELDT. It should also be noted that a new standard setting applying to the 2006–07 CELDT will make the *early advanced* performance standard more difficult to attain (discussed further below).

RETURN to TABLE of CONTENTS

Student Level Progress Patterns

While the Figure 5 shows overall cross-sectional performance distributions and the CELDT test years utilized for AMAO calculations, other data are needed to shed light on how well ELL students are meeting AMAO 1 growth targets and AMAO 2 proficiency targets by CELDT level. Figure 6 displays the percentage of English learners meeting AMAO 1 growth targets by CELDT proficiency level for 2004–05 and 2005–06. As can be seen, about two-thirds of ELLs at *beginning* and *early intermediate* proficiency levels met their growth targets in both years, and these ELLs accounted for 30% to 32% of all ELLs in the AMAO 1 cohort, depending upon the year. Notably, less than half of the ELLs at *intermediate* met their growth target, and these students also constitute the largest proportion of the cohort in each year examined (36% to 38% of the AMAO 1 cohort, depending upon the year). Clearly, as was the case in our baseline cohort data, it is more difficult for these ELL students to advance to the next language proficiency level in a single year. Moreover, since only two years of matched-score CELDT data are available at any one time, there is currently no way to know *how long* ELL students who advanced from *intermediate* had been at that level. It is possible that many had been there for more than one year.[11]

Figure 6. ELLs meeting AMAO 1 growth targets by CELDT level, 2004–05 and 2005–06.

	2004–05		2005–06	
Growth Target (From prior to current year CELDT)	% ELL meeting target	% AMAO 1 cohort N = 1,314,616	% ELL meeting target	% AMAO 1 cohort N = 1,292,977
Beginning to *Early intermediate*	65	14	65	14
Early intermediate to *Intermediate*	67	18	66	16
Intermediate to *Early adv/Adv.*	49	38	45	36
Early adv/Adv. to *Eng. prof.* level	51	3	51	3
Maintain *English proficient level*	80	28	81	31

Figure 7. ELLs meeting AMAO 2 growth targets by CELDT level, 2004–05 and 2005–06.

	2004–05		2005–06	
AMAO 2 cohort inclusion criteria (based on prior-year CELDT level)	% ELL meeting target	% AMAO 2 cohort N = 728,562	% ELL meeting target	% AMAO 2 cohort[12] N = 613,185
Intermediate	44	69	41	75
Early adv/Adv., not *Eng. proficient*	51	5	51	7
Beg./Early int., 4 or more years in U.S. schools	10	23	9	13
Beg./Early int. < 4 years in U.S. schools, reaching *Eng. prof.* level	100*	4	100*	5

**100% by definition, as category includes only those beginning/early intermediate ELLs in U.S. schools less than four years that reach the English-proficient level in the current year.*

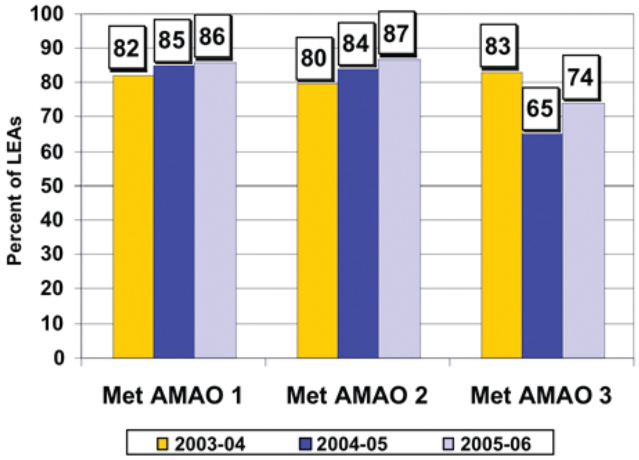
Also, a sizable proportion (around 80%) of ELLs at the *English proficient* level on CELDT in the prior year—who did not meet the additional criteria for reclassification and so were retested on CELDT the following fall—were able to maintain the equivalent level of language proficiency. This group constituted the second largest proportion of the cohort in each year examined, and may signal a need for educators to support ELLs in satisfying the additional criteria that may be preventing them from being reclassified out of the ELL category. Finally, just over half of the 3% of the cohort for each year examined at the *early advanced* or *advanced* CELDT levels were able to bring all of

[11] California is in the process of implementing a statewide pupil database that will make multiyear longitudinal analyses possible.

[12] AMAO 2 cohort number decreases in 2005-06 due to a change in the required U.S. enrollment variable, which previously required spring of first enrollment year, and changed to month/day/year of first enrollment, resulting in more missing cases.

RETURN to TABLE of CONTENTS

Figure 8. California LEAs meeting individual AMAOs, 2003–04 to 2005–06.



their subskill scores up to *intermediate* and meet the *English proficient* level.

With respect to AMAO 2, Figure 7 displays the percentage of English learners meeting AMAO 2 proficiency targets by CELDT proficiency level for the two most recent years of AMAO 2 calculations. As can be seen, ELL students with a prior CELDT level of *intermediate* constitute the largest proportion of those in the AMAO 2 cohort—from 69% to 75%, depending upon the year. Fewer than half of these students reach the *English proficient* level in the next year. Of arguably greater concern are those ELLs in AMAO 2 cohort that are at the *beginning* or *early intermediate* levels and in U.S. schools for four or more years. They constitute almost one quarter (23%) of the cohort in 2004–05, and only 10% of them reach the *English proficient* level on the next CELDT administration. Notably, the proportion of these students decreases in 2005–06 to 13%, yet only 9% of the students attain the *English proficient* level the next year.[13] As seen in AMAO 1 analyses above, just over half of the relatively small percentage of ELLs in the AMAO 2 cohort at *early advanced* or *advanced* CELDT proficiency levels (5% to 7% depending upon year) attained the *English proficient* level on taking the CELDT again. Finally, an equally small percentage (4% to 5%) of ELLs in this cohort previously scoring at the *beginning* or *early intermediate* levels and in the U.S. schools less than four years jumped to the *English proficient* level upon taking the CELDT the following year.

AMAO Results at the LEA Level

Figure 8 shows for each of the three last years the percentage of Title III LEAs that met each AMAO target. Recall that for AMAOs 1 and 2 the baseline performance level was set at the 20th percentile of the LEA distribution (i.e. where 80% of LEAs met the target and 20% did not). There is a steady increase in the proportion of LEAs meeting individual targets for these two AMAOs. That is, greater proportions of California's LEAs (from 82% to 86%) are meeting *increasing* targets for the percentage of ELL students making progress in learning English, and those attaining the English-proficient level (from 80% to 87%), as measured by CELDT. However, a different pattern is evident for AMAO 3 in which the percentage of LEAs meeting the AYP AMAO dropped from 83% to 65%, then recovered to 74%, as measured by the California Standards Tests and the California High School Exit Exam in English language arts and math.[14] The decrease in 2004–05 can be attributed in part to the stairstep increase in the AYP performance target that occurred between 2004 and 2005 (see Figure 4). Also, as mentioned above, the status bar nature of AMAO 3 defines progress only as the percentage of ELL subgroup who scored proficient or above in academic assessments of English language arts and math, with no consideration given to the annual progress of students occurring below the proficient threshold.

Given that Title III requires LEAs to meet *all three* AMAOs each year, Figure 9 displays the percentage of Title III LEAs that met both AMAO 1 and AMAO 2 (using CELDT only), as well as those meeting all three AMAOs (using CELDT and academic measures). As can be seen, the proportion of LEAs meeting the first two AMAOs has increased steadily over the past three years from 77% to 84%. However, once the AYP AMAO is included, the proportion of LEAs meeting all three AMAOs drops notably, and mimics the sawtooth effect seen for AMAO 3 in Figure 8.

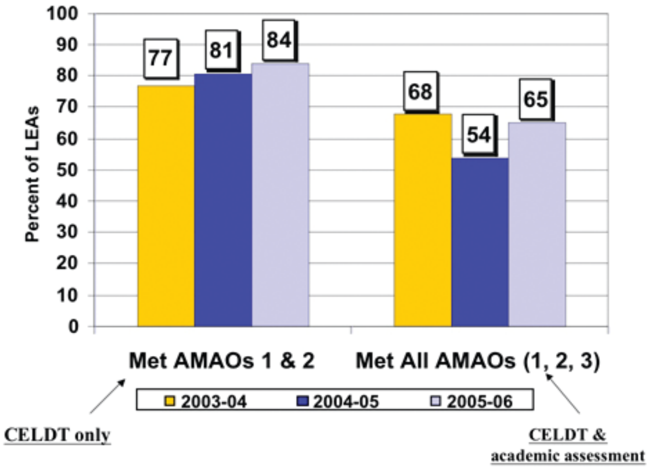
As can be clearly seen, the AYP AMAO has an enormous effect on outcomes in California's Title III accountability system. In fact, 182 LEAs in 2005, and an additional 103 LEAs in 2006, missed one or more AMAOs for two

[13] Since the proportion of *intermediate* level students does not increase from 2004 to 2005, the drop in proportion of this time-defined group is likely due to the enrollment variable change described in the previous footnote.

[14] In addition to CST and CAHSEE, the California Alternate Performance Assessment (CAPA) results are also used in Title I AYP and Title III AMAO 3. See: http://www.cde.ca.gov/ta/ac/ay/documents/infoguide06.pdf.

RETURN to TABLE of CONTENTS

Figure 9. California LEAs meeting two or more AMAOs, 2003–04 to 2005–06.



years and therefore have been identified under this accountability system as needing to improve services for their English learners. Many of these include the largest ELL-enrolling districts in the state, and the large majority of these have been identified due to AMAO 3, the AYP AMAO.

Conclusion

Assessment and accountability researchers remind us that the value of an accountability system—particularly one with high stakes—should be judged, among other things, by the degree to which it is technically defensible, comprehensible, based on ambitious yet reasonable expectations, and useful to educators and the public (Baker et al., 2002; Linn, 2003; Linn et al., 2002). The Title III accountability system established and in place in California since 2003, though far from perfect, has been widely understood and accepted by educators and policymakers. This is especially significant as the education of English language learners has long been a politically charged subject in California, and topics such as expected annual progress and time to English proficiency can generate significant debate. We believe contention was avoided and consensus built due in large part to two key components of the development process: First, the AMAO development team was able to model and recommend options to policymakers based on empirical data from the students and districts to be affected by the accountability system. That is, the AMAO team was able to demonstrate what outcomes were likely for this population under the different options considered. This grounded the policy discussion and gave policymakers a clear sense of the implications of their decisions. Second, the AMAO team provided opportunities for educators in the field to give input at key points in the development process. This fostered understanding and trust, and yielded insights that were ultimately incorporated into the proposals for policymakers. For example, based on feedback from the field, the draft AMAO 2 cohort definition was modified to "give credit" to LEAs for their ELLs at *beginning* and *early intermediate* levels reaching the *English proficient* level in less than four years. As a result of using empirical data and an inclusive process, diverse stakeholder groups publicly supported the AMAO team's recommendations, and the State Board unanimously approved their adoption.

Since its introduction, California's Title III accountability system has evolved as new needs emerge and conditions change. For example, educators at the LEA level requested that the state provide **school-level** AMAO 1 and AMAO 2 results, in order to facilitate discussions at the local level of what each school was contributing to the overall district outcome. As a result, the state provides district level outcome reports, as well as school results.[15] The AMAO performance patterns discussed above have been regularly shared in multiple venues with educators across the state in order to highlight the need to put faces and names to these numbers at the district- and school-site levels. School districts now appear more likely to know and focus on those ELLs who are not progressing. Indeed, in large part due to these AMAOs, many Title III LEAs are locally monitoring and analyzing their students' longitudinal CELDT performance—as well as ELL progress during the year via diagnostic assessments—in order to illuminate patterns of instructional need for students and professional development priorities for teachers. This kind of multi-year longitudinal monitoring of ELL progress has long been recommended for local and state levels (NRC, 2000). CDE is continuing to monitor outcomes on the ELP AMAOs for evidence of possible disparate impact by type of LEA (elementary, secondary, unified, etc.). Moreover, after resetting performance standards for the 2006–07 CELDT, the state is also recalibrating 2005–06 CELDT results on those new performance standards before calculating its 2006–07 AMAOs to ensure test-result comparability.

[15] See http://www.cde.ca.gov/sp/el/t3/acct.asp for AMAO information guides, and district and school reports.

RETURN to TABLE of CONTENTS

As shown above, importing Title I AYP into the Title III accountability system—which is specifically required by NCLB—has had a major impact on the identification of LEAs in California. From a local accountability standpoint, one clear message emerging for districts is that they need to attend more carefully to the meaningful access that ELL students have to grade-level academic core content via appropriate instruction. Many ELLs have been in the California school system for many years, and former ELLs meeting local reclassification criteria are largely included in this subgroup calculation as well. As a result, more California educators are focusing on those ELL students at the basic level of CST-ELA performance in the belief that these students have the best chance to achieve academic proficiency and improve subgroup results. From a consequential validity standpoint, numerous issues emerge from incorporating Title I AYP into Title III accountability. The threats to validity in using academic achievement assessments in English without careful regard to students' English language proficiency, time in U.S. schools, language of instruction, etc., are well documented (e.g., Abedi, 2004; NRC, 1999; NRC, 2000). In addition, Title I's use of a single-year status bar approach does not credit progress students make throughout the performance range below the proficient level. These issues are being openly discussed nationally as NCLB approaches reauthorization, and initiatives such as the U.S. Department of Education's LEP Partnership and Growth Model Pilot Project are specific examples of attempts to address dilemmas generated by NCLB in ways that do not undermine its intent.

There may be opportunities to learn from the strengths of the *de facto* growth model approach encouraged by Title III, which allows ambitious yet defensible targets to be set empirically for progress under AMAO 1, and eventual attainment of English language proficiency under AMAO 2. Specifically, the conceptual framework underlying and interrelating these two AMAOs, and empirical methods like those used in California to develop them, could offer a helpful (and hopeful) approach as state and federal policymakers attempt to adjust Title I policy in order to enhance equity and discourage cynicism or gaming of the accountability system.

References

Abedi, J. (2004). The No Child Left Behind Act and English-language learners: Assessment and accountability issues. *Educational Researcher*, *33* (1), 4–14.

Baker, E., Linn, R., Herman, J., & Koretz, D. (2002). *Standards for educational accountability systems*. CRESST Policy Brief 5 (Winter). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Crawford, J. (2002). Programs for English Language Learners (Title III). In *ESEA Implementation Guide*. Alexandria, VA: Title I Report Publications.

De Ávila, E. (1997). *Setting expected gains for non and limited English proficient students*. NCELA Resource Collection Series No. 8. Washington D.C.: National Clearinghouse for English Language Acquisition.

George, C., Linquanti, R. & Mayer, J. (2004). *Using California's English Language Development Test to implement Title III: Challenges faced, lessons learned*. Symposium paper presented at Annual Meeting of the American Educational Research Association, San Diego, CA.

Gottlieb, M., & Boals, T. (2006). *Using ACCESS for ELLs data at the state level: Considerations in reconfiguring cohorts, resetting annual measurable achievement objectives (AMAOs), and redefining exit criteria for language support programs serving English language learners.* WIDA Technical Report #3. Madison: WIDA Consortium, Wisconsin Center for Educational Research, University of Wisconsin.

Government Accountability Office (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency*. Report GAO-06-815 (July). Washington, DC: Author.

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

Linn, R. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, *32* (7), 3–13.

RETURN to TABLE of CONTENTS

Linn, R., Baker, E., & Betebenner, D. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31* (6), 3–16.

Linquanti, R. (April, 2004). *Assessing English-language proficiency under Title III: Policy issues and options*. Symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners*. Policy Report 2001–1. Santa Barbara: University Of California Linguistic Minority Research Institute.

National Research Council (NRC). (2000). *Testing English language learners in U.S. schools*. Kenji Hakuta and Alexandra Beatty, Eds. Washington, DC: National Academy Press.

National Research Council (NRC). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Jay P. Heubert and Robert M. Hauser, Eds. Washington, DC: National Academy Press.

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

World-Class Instructional Design and Assessment (WIDA) Consortium. (2006). Composite score growth profile based on years 1 & 2 ACCESS for ELLs data. WIDA Consortium draft recommendations. Madison: Author.

Zehr, M. (2006). New era for testing English-learners begins. *Education Week*, *25* (42).

RETURN to TABLE of CONTENTS

Chapter 8

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

Chapter 9

Summary and Recommendations

Jamal Abedi

Fairness demands increased efforts towards improving the quality of assessments for English language learners (ELLs), especially in light of the performance gap between ELLs and their native-English-speaking peers. Inadequacies in the assessment and instruction of ELL students may partly explain such gaps, so fair assessment is a priority. These issues raise important equity considerations, especially as the population of ELL students increases rapidly.

One of the greatest influences on ELL students' academic careers is their level of proficiency in English, given that they are primarily instructed and assessed in English. Assessments of English language proficiency based on questionable measures may cause grave academic consequences. ELL students who are inappropriately assessed may be misclassified with respect to their level of proficiency in English and may receive inappropriate instruction. They may even be misclassified as students with learning disabilities, which may greatly impact their academic career (see, for example, Abedi, 2006; Artiles, Rueda, Salazar, & Higareda, 2005).

Due to the importance of adequately assessing a student's level of English language proficiency (ELP), the NCLB legislation requires that schools receiving Title I funding assess ELL students using reliable and valid measures. With this mandate, the legislation plays an important role in bringing the need for English language assessment to the forefront of education accountability.

In addition to emphasizing the need for ELP assessment, the NCLB Title III legislation provides a set of guidelines for constructing ELP assessments that render reliable and valid estimates of a student's level of English proficiency. These guidelines provide specific tools to help the measurement community be more vigilant of ELL assessment needs and to be better prepared for assessing ELLs. In chapter 1 of this report we discussed some shortcomings of ELP assessments developed before the implementation of NCLB. Many of the pre-NCLB assessments were not based on an operationally defined concept of English proficiency, had limited academic content coverage, were not consistent with states' content standards, and had psychometric flaws.

NCLB Title III contributes greatly to improving the quality of ELP assessments in many different ways. By making such assessments a requirement for ELL students, NCLB encourages the education community to pay greater

RETURN to TABLE of CONTENTS

attention to this area. The NCLB description of ELP assessment requirements also help define ELP assessments more operationally. For example, NCLB requires ELP assessments to include four domains (reading, writing, speaking, and listening), measure student's academic English proficiency, and be aligned with the states' ELP standards—as well as content standards—across three academic topic areas and one non-academic topic area related to school environment. By introducing the concept of academic English and academic content into ELP assessment, NCLB requires states to measure ELL students' academic success more directly. As noted by more than one author in this report, content standards are more evident in the classroom instruction of ELLs, partly because the tests reflect them.

These are significant milestones in the history of ELP assessment. By assessing academic language proficiency, states more thoroughly address language needs related to academic success. Alignment of ELP assessment content with the states' ELP content standards provides more authentic assessments that are relevant to students' academic needs. Other requirements, such as evidence on the reliability and validity of assessments and introducing the testing across grades K through 12, also contribute to improved assessment of English language proficiency.

The four consortia that carried out the challenging task of developing post-NCLB assessments carefully followed the ELP assessment requirements under NCLB and produced quality work. These ELP assessments include items in the four domains of reading, writing, listening and speaking and provide outcome measures for each as well as scores for comprehension (listening and reading) and for overall performance. The consortia also conducted standard setting studies to set achievement levels in several categories including *basic*, *proficient* and *above proficient* for all four domains. ELP tests were developed for four grade clusters (kindergarten through grade 2, grades 3 through 5, 6 through 8, and 9 through 12). The four chapters contributed by the four consortia briefly discussed such efforts in producing the ELP assessments that many states are currently using. The newly developed assessments underwent extensive pilot testing and field testing on large and representative samples of students. The content and psychometric properties of the individual items as well as the total tests were carefully examined and improvements were made.

The newly developed ELP assessments are based on the theoretical framework of second language acquisition and other principles in the field of linguistics (Bauman, Boals, Cranley, Gottlieb, and Kenyon, 2007; see also, Cummins, 1981 and Thomas, 2002). The assessments were also informed "by second-language development theory of communicative competence which posits that ELP tests should measure communicative and participatory language in the context of the classroom and that they should be age/grade appropriate" (Lara, et al., 2007, p. 48). In addition to introducing such important principles into the ELP test development process, the creators of new ELP assessments designed sets of standards that are common across the participating states using sound alignment methodologies (see, for example, Bauman, et al.; Lara, et al.; Rebarber, et al., 2007). Creation of the final products was informed by the results of pilot and field studies. The field studies were based on large representative samples of students in each of the grade clusters. Detailed item statistics were obtained and examined for any significant problems, and then achievement levels were set using sound methodologies (Bauman, et al.). The technical details of these activities are reported in the technical manuals of these assessments.

The innovation, collaboration, process, and amount of teacher participation in developing the consortia assessments set a model for future test developers. Rather than focusing on a particular grade or age level, these assessments cover kindergarten through grade 12, using a grade-cluster approach that grouped test instruments into four clusters: K–2, 3–5, 6–8, and 9–12. There were substantial numbers of common items built into the assessments to facilitate vertical scaling across the clusters (Lara, et al., 2007; Bauman, et al., 2007; Mathews, 2007; and Rebarber, et al., 2007). In addition, many of the consortia used test development methodologies that facilitate understanding of the developmental nature of the assessments. For example, the validation study of ELDA was based on the latent-class methodology in which items were studied in terms of ELP development (Lara, et al). Similarly, WIDA (Bauman, et al.) conducted developmental level ratings of items in which educators experienced in teaching English-language learners and/or language testing were asked to assign items with a performance-level designation that best identified the language proficiency level necessary to answer each item correctly. In addition, the use of graphics, spoken language prompts and on-the-spot scoring are found among the development reports. Bias review, an important process in ELL assessment development, was performed by all the consortia. While these exemplary efforts have established a

RETURN to TABLE of CONTENTS

solid foundation for ELP assessments, we believe there are still issues to resolve. It will take more attention and work to bring the new assessments to the level of providing reliable and valid ELP outcome measures.

Improving the New Generation of ELP Assessments

In this section, we will elaborate on the essential considerations for cultivating instruments into valid and reliable ELP assessments. Although addressed somewhat in the consortia chapters, we need to examine: (1) English language proficiency (ELP) standards, (2) standard setting for ELP assessments, (3) *dimensionality* (the relationship among the parts of a test to its whole), (4) pilot and/or field-test findings, (5) the *baseline* scores used in reporting progress, and (6) the concept of *academic language*. In this chapter, we will briefly discuss these issues and provide recommendations based on both our review of the literature on existing assessments (August, Francis, Hsu & Snow, 2006) and on our knowledge of the newly developed ELP assessments.

1. ELP Standards

As mentioned earlier in this report, NCLB requires states to first develop English language proficiency (ELP) standards suitable for ELLs. Then, based on these standards, states implement a single, reliable and valid English language proficiency assessment that annually measures reading, writing, listening, speaking, and comprehension and is aligned to state ELP standards.

However, this raises the question of which ELP standards and from which of the participating states? Many states participating in one of the four consortia did not have a set of well-defined ELP content standards at the beginning phase of the implementation of the Title III assessments (e.g., Fast, Ferrara & Conrad, 2004). Even if all participating states in a consortium had well-established ELP content standards at the start of the ELP development process, which standards should all participating states use to develop a common assessment? Are there a set of common ELP standards across the participating states in the consortium? If there are standards in common, do they have the same level of importance for all the participating states? If not, how should the alignment of ELP assessments with state standards be addressed?

2. Standard Setting for New English Language Proficiency Tests

Different states use different approaches of standard setting (Texas Education Agency, 2002). Even within the same approach for standard setting, the results may vary greatly across states depending on factors such as the educational background and training of the judges involved in the standard setting process. For example, there are reports of inconsistencies between the achievement level outcomes produced by different techniques (e.g., Impara & Plake, 1997; Jaeger, 1989; Kiplinger, 1997). As a source of inconsistency between states' assessment results, Musick (2000) refers to different sets of standards for student learning. According to Musick, some states may have lower performance standards for student achievement than others. Musick reported substantial differences between states in the percent of students meeting state performance standards; for example, according to Musick, the percentage of students meeting state standards in grade 8 math ranged between 13% in one state and 84% in another state (Musick, 2000, Table 1, p. 4). When the same two states were compared on their NAEP performance scores, Musick indicated that the state with the lowest percentage meeting state standards actually scored higher on NAEP than the state with the highest percentage.

Loomis (2001) evaluated the outcome of different standard setting procedures by comparing teachers' judgments of student performance to the empirical classification of student performance (such as with the contrasting group approach) and to the performance represented in test booklets (such as the Bookmark and Modified Angoff methods). He concluded that there was no certain way to verify the validity of the cut scores. Jaeger (1989) groups several different standard setting methods under two major categories: test-centered models and examinee-centered continuum models. The test-centered models include Angoff, Modified Angoff, Ebel's Procedure, Jaeger's Procedure, and Nedelshy's Procedure. Under the examinee-centered category, Jaeger lists the Borderline-group procedure and Contrasting-groups procedure.

Jaeger (1989) warned that different procedures may produce very different results. Jaeger summarized 12 different studies that compared the results of using different standard setting methods with the same tests under similar conditions (Table 14.1, pp. 498-499). A total of 32 contrasts among the different methods were reported. For each

RETURN to TABLE of CONTENTS

contrast, the ratio of the largest test standard to the smallest test standard was computed. Based on these results, Jaeger (1989) indicated that "at best, the ratio of the largest recommended standard to the smallest recommended standard was 1.00," indicating identical standards resulting from different methods. At worst, however, "the recommended standard resulting from one method was 42 times as large as that resulting from another method" (p. 500). Based on the findings of this study along with recommendations of others (e.g., Hambleton, 1980; Koffler, 1980; Shepard, 1980; and Shepard, Camilli, & Williams, 1984), Jaeger recommended that the results from several methods must be used in setting standards.

Based on this brief discussion it is clear that there may not be a sense of comparability across states on their ELP standard setting outcomes. States do not often apply different standard setting approaches to examine possible discrepancies between the outcomes of these approaches. More importantly, we are not sure how replications of standard setting would produce similar results within a particular state.

In addition to the sources of inconsistencies due to the use of different standard setting approaches, other factors may introduce bias into the process. For example, in deciding the number of performance levels, the consortia were faced with a dilemma. Fewer cut-points may require fewer items; thus, shorter tests. A greater number of performance levels provides the opportunity for more subtle distinctions between students' performance, but requires a greater number of items and longer tests (Fast, et al., 2004). Typically, five performance levels were used in the newly developed ELP assessments. The performance level descriptors (PLD) were slightly different across the different tests developed by the four consortia but *Level 1* usually refers to *no* or *very low proficiency* in English and Level 5 represents *high proficiency*. Similarly, different states set their criteria for reclassification of ELL students from *limited English proficient* (LEP) to *fluent English proficient* (FEP) at different proficiency levels but ELL students are typically reclassified from ELP to FEP at ELP performance *Level 4* or above.

There are other issues concerning standard setting for the ELP assessments as well. Among these issues are inconsistencies between achievement levels set for the different domains of reading, writing, speaking, and listening. When achievement levels are set separately for each of the domains, then discrepancies between such levels across the domains could make interpretation of the results difficult. For example, many students can be classified as *proficient* or above in one domain but may be classified as *below proficient* in other domains. How can such issues be resolved? (see for example, Bunch, 2006). Should achievement levels be set at the whole test level? If so, then how should the total test score be obtained? This raises a whole new set of issues concerning dimensionality, the topic of our next section.

3. Dimensionality Issues

NCLB Title III requires states to measure the annual growth of students' English language development in reading, listening, writing, and speaking—and comprehension. In addition to the scores from each of these four domains, composite scores of all domains as well as groups of subscales are used. The overall composite of the four subscales is commonly used by states. However, other composite scores based on some subscales are also used by the member states of the consortia. For example, ACCESS for ELLs' "[r]eports include four weighted composite proficiency scores: an Overall composite score reflecting all domains, an Oral Language composite score (listening and speaking), a Literacy composite score (reading and writing), and a Comprehension composite score (listening and reading)" (Bauman, et al., 2007, p. 90). Sometimes these composites are based on unequally weighted subscale scores. For example, based on the input from the member states, the WIDA consortium (ACCESS for ELLs®) computed the overall composite as 15% listening, 15% speaking, 35% reading, and 35% writing and the comprehension composite as 30% listening and 70% reading. Similar policies have been adopted by other consortia of states.

However, to create these composite scores, it is important to know how these different subscales are correlated and whether they measure a single construct, i.e. English language proficiency, or whether they measure four different constructs, namely reading, writing, listening and speaking. If they are highly correlated, the decision regarding combining the different subscales as well as the weightings would be more understandable than when the subscales are not highly correlated. Therefore, the issue of dimensionality needs to be addressed prior to such decisions. It is also important to provide evidence to justify the use of particular weights used to create composite scores. How their weights are selected and whether scores on one

RETURN to TABLE of CONTENTS

particular subscale are weighted higher than others is important to consider. There can be a big difference between decisions based on the views of state policy makers and those based on solid empirical evidence.

To begin with, researchers should ask, "should the four domains be considered as four separate subscales/dimensions or should they be considered as a single latent trait that encompasses all four domains?" There are different models and different views on this choice. The number of constructs being measured seriously affects reporting and interpretation of scores. If the four domains are measuring a single construct (i.e., the overall English language proficiency latent variable) then scores from the four domains can be combined and a single score can be used for reporting AMAOs and for classification purposes. On the other hand, if each domain has a unique contribution to the ELP construct, how can a total score be obtained and interpreted? Different models for combining subscale scores are suggested in the literature (see, for example, Abedi, 2004; Sawaki, Stricker, & Oranje, 2007). In a factor analytic approach, when a single ELP construct is postulated, the common variance shared across the four domains is used and a latent variable of ELP is computed. This requires high or near perfect correlations between scores of the four domains. If subscales contain some specific variance in addition to the overall ELP factor, then among the two most commonly used options (compensatory versus conjunctive models), which one would be preferable? In the compensatory model, a low score in one domain may be compensated by high scores on another domain. As Abedi (2004) elaborated, the preferred model for NCLB is the conjunctive model in which students should score at the proficient level in each of the four domains to pass AMAO requirements.

4. Pilot and/or Field-test Findings in Test Development

A major strength of the assessments developed by the consortia was the well-designed validation studies incorporated into the development process. Many different approaches in validation of the ELP assessments were utilized. These approaches include latent-class analyses, criterion-related approach using both concurrent and predictive approach as well as a Multi-Trait/Multi-Method approach within a structural-equation modeling framework, and content validation through alignment to ELP content standards and construct validations using the confirmatory factor analytic approach (e.g., multiple group confirmatory factor analyses).

In many of these analyses, the developmental nature of these assessments was considered. For example, the latent-class model was applied in validating ELDA (Lara, et al., 2007), as indicated in the discussion of ELDA (Chapter 4):

> The framework for these latent class analyses is the theoretical view of English language development in which an English language learner passes through multiple stages of development, from pre-production to advanced fluency, in each of four major modes—listening, speaking, reading and writing—that are reflected in the four domains assessed by ELDA. (pp. 54–55)

A major problem in estimating the validity of ELP assessments through the criterion-related or construct approach are issues concerning content and psychometric characteristics of the criteria used for validation of ELP measures. As indicated earlier, reviewers of the pre-NCLB assessments expressed concern over some of the tests' soundness and validity (Del Vecchio & Guerrero, 1995; Zehler et al., 1994). A low correlation between the newly developed ELP assessments and the older assessments might be expected since the content and structure of these assessments might be quite different.

Bauman, et al. (2007) found correlations between the four domains of a new ELP assessment (reading, writing, listening, and speaking) and four existing ELP assessments (IPT, LAS, LPTS, and MAC II) to range between .468 and .765 with an average of .604. While this correlation is considered relatively high, it explains only 36% of the variance between the ACCESS for ELLs test and the existing ELP tests used as the criterion variables. Once again, there are many factors that could explain the lack of a strong correlation between the newly developed ELP assessments and the existing ELP assessments. Among the most important sources contributing to a low correlation is the low content and psychometric comparability between the two sets of assessments.

RETURN to TABLE of CONTENTS

5. Baseline Scores for the NCLB Title III Assessment Reporting

As the implementation phase of NCLB Title III began around 2002, efforts to develop new ELP assessments based on the NCLB requirements began as well. It took over three years for most of the consortia's assessments to become fully developed and field tested. In 2002, there were many existing ELP assessments on the market—a majority of which; however, did not meet the NCLB Title III assessment requirements. Since the newly developed ELP assessments were not available at the start of NCLB implementation, states had no other choice but to use whatever existing ELP assessment they found relevant. This situation obviously introduced flaws into the reporting of ELP progress, one of which was that subsequent tests might not be comparable with the tests they replaced. Determining reasonable annual growth expectations in English; operationally defining the English proficient level; and setting baselines and annual growth targets for local education agencies are among the problems that the states faced (George, Linquanti & Mayer, 2004; Gottlieb & Boals, 2006; Linquanti, 2004).

Now that many states have access to the newly developed ELL assessments that meet the NCLB requirements, they are faced with the quandary of linking baseline results based on *off-the-shelf* ELP assessment tests with the results from their new ELP assessments. The problem is not limited to ELP assessment content, i.e., not having access to assessment outcomes in the four domains (reading, writing, speaking and listening), other problems and issues exist that make such comparisons a real challenge. For example, as indicated earlier, many of the existing assessments at the start of NCLB Title III implementation were based on different theoretical emphases prevalent at the time of test development. In addition, they were not aligned with state's ELP content standards and did not reflect the importance of academic language development. Therefore, even a high statistical correlation between ELP assessments used as the baseline and the new ELP assessment would not be enough to establish a strong link between the two assessments.

The WIDA consortium conducted a study that may provide evidence to support the use of pre-NCLB assessment measures as the baseline. This bridge study was conducted with a sample of 4,985 students enrolled in grades K through 12 from selected districts in Illinois and Rhode Island. Students in this study took both the ACCESS for ELLs® and one of four older English language proficiency tests: *Language Assessment Scales* (LAS), the *IDEA Proficiency Test* (IPT), the *Language Proficiency Test Series* (LPTS), and the *Revised Maculaitis II* (MAC II). The purpose of this study was to predict performances on ACCESS for ELLs® from performances on the older tests using a linear regression procedure.

One can argue that if ELP assessments claim to measure students' levels of English proficiency, there must be a high intercorrelation between those assessments —whether they are pre- or post-NCLB. However, as Bauman, et al. (2007) elaborated, one may not expect high-level relationships between the pre- and post-NCLB assessments since they are very different in many different aspects, including content, construct and psychometric characteristics. The findings of the bridge study by Bauman, et al. suggest a moderate-to-high-level relationship between the four domains of ELP assessment (reading, writing, listening, and speaking) and four existing ELP assessments (IPT, LAS, LPTS, and MAC II).

6. The Concept of Academic Language

Among the NCLB instructions provided to states for developing reliable and valid ELP assessments is to incorporate the concept of academic English into the process. The major goals of NCLB Title III are:

> to help ensure that limited English proficient (LEP) children attain English proficiency, develop high levels of academic competence in English, and meet the same challenging state academic content and student academic achievement standards that all children are expected to meet. (Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students, 2003, p. 5)

For example, in response to the question: "B-5. Why must English language proficiency standards be linked to academic standards?" the U.S. Department of Education indicated that:

> The statute requires English language proficiency standards to be linked to state academic content and achievement standards in reading or language arts and in mathematics beginning in the school year 2002-2003. This is required in order to ensure that LEP students can attain proficiency in both English language and in reading/language arts, math and science. English language proficiency standards should

RETURN to TABLE of CONTENTS

> also be linked to the state academic standards in science beginning in the school year 2005-2006. (Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students, 2003, p. 10)

Clearly, the focus of the ELP assessment mandate is academic English. Many of the newly developed measures of ELP; therefore, are based on the need to test academic English which in turn facilitates learning content knowledge across the following three academic topic areas: English/Language Arts, Math, Science and Technology, and Social Studies as well as one non-academic topic area related to the school environment (Fast, et al., 2004).

However, there have been controversies over what is covered under the concept of academic English. Should the ELP assessment include the language in the above content areas or cover the language that facilitates learning of the contents? Fast, et al. (2004) clarify this issue by indicating that ELP assessments "are not tests of academic content, in other words, no external or prior content-related knowledge is required to respond to test questions." (p. 2). That is, eventually, ELL students should be able to demonstrate the full level of English proficiency that enables them to successfully function within the appropriate grade level.

Even at this level of clarity, who decides how academic English proficiency should be captured within the ELP assessments? How should we evaluate the content and psychometric properties of ELP assessments that test academic English proficiency? How can one recognize an assessment that is not testing academic English?

We believe this is an area that needs attention from experts in a variety of disciplines. Experts in the field of linguistics with knowledge and experience in academic language, along with content and measurement experts, should join in the effort to operationally define academic language and provide guidelines for test item writers who are assigned to ELP test development. It is also important to include teachers, bilingual coordinators and state personnel working with ELP assessment experts. Meanwhile, it might benefit states to review their current ELP tests and evaluate the test items in terms of academic English content.

The Comparability of ELP Assessments

Everyone who followed the NCLB Title III instructions and developed ELP assessments has presented assessments with high levels of technical quality. For example, the assessments are aligned with the ELP standards of member states and content area standards, they test across a range of age/grade levels, they test in the four language domains, and they assess academic English skills. Furthermore, the test items went through a rigorous validation process in which the items were examined for any sign of bias or technical problems. Once again, a review of the test development process provides evidence of the high quality of these assessments.

However, the consortia and commercial assessments were developed separately with no interaction between the test developers across the projects. Therefore, as of this point, there is not enough evidence to make any judgment about cross-validity or cross-comparability of these assessments.

With this in mind, we invited the head project officers for the four ELP consortia of states, along with assessment directors from the major publishers involved in post-NCLB development of ELP assessments, to a joint effort to discuss their work and share their ideas on three different occasions: (1) a pre-session at the 2006 Large-Scale Assessment Conference in San Francisco, (2) a professional training session at the 2007 American Educational Research Association in Chicago, and (3) a pre-session at the 2007 Large-Scale Assessment Conference in Nashville. The outcome of these joint meetings is this report which presents an overall picture of what is happening in terms of ELP assessment across the nation. However, so far, no real data on the performance of these assessments for comparison purposes have been exchanged.

While it would be difficult to plan efforts to link these assessments, it is imperative to compare the content and technical aspects of these assessments. We understand that states are independent in their decisions for selecting and using assessments, but we also believe that comparisons of these data would provide a wealth of information regarding the validity of assessments that would be impossible to obtain from any of the existing sources.

Below is a set of recommendations based on our understanding of the field as well as information provided to us in the process of compiling this report. Once again, we hope our presentations along with our recommendations will start a national dialog for improving the quality of ELP assessments. This is a necessary step in providing better instruction, assessment and accountability systems for ELL students.

RETURN to TABLE of CONTENTS

Recommendations

Based on the information presented in different chapters of this report and based on the review of existing literature on the English language proficiency assessments, below are some recommendations. These recommendations may help states to improve their existing Title III assessments and plan for more reliable and valid ELP assessments in the future.

- Use multiple methods for setting standards and defining cut scores for achievement levels.
- Examine the comparability of the assessment used to establish the baseline with the newly adopted ELP assessment. These comparisons should be done in both content and psychometrics of the assessments. If there is not a strong link between the two assessments both in term of content and psychometric characteristics, then use caution in making firm judgments about the growth of students' English proficiency.
- Examine the content of the state-adopted ELP assessment and align the content with the state ELP content standards. This is important, since the ELP consortia's assessments are not completely based on the ELP standards of any one state.
- Examine the pattern of possible differential performance of ELL students on the ELP assessments to make sure that the ELP assessment items do not differentially or unfairly perform across the subgroups within the ELL population.
- Use multiple criteria for assessing ELL students' level of English proficiency, particularly with high-stakes decisions such as classification or re-classification of students.
- Use ELP assessment results along with other sources to make informed decisions about ELL student participation in Title I assessment, as the literature clearly suggests that assessments that are constructed for native speakers of English may not provide valid outcomes to ELL students at the lower levels of English proficiency.
- Train staff with high levels of knowledge and experience in measurement in order to constantly review and monitor assessment issues, particularly in the area of English proficiency. As a part of their contract, test publishers may provide states with the needed technical information. In addition, states should have an independent evaluation capacity to examine the quality of state assessments.
- Incorporate a major measurement research component into programs that can be supervised and run with professionally trained staff and consult with standards-based recommended guidelines for improving the quality of the ELP assessments (see, for example, Robeinowitz and Soto, 2006). Once again, states must always reserve the right to examine the validity of their assessments and conduct analyses independent of what the test publishers/developers provide, to bring another level of confidence into their high-stakes assessments.

Further, it is imperative that states actively pursue research and development in maintaining the quality of ELP assessments. Conducting validity studies and examining test items for any indication of threats to their authenticity over time will help assure the quality assessment of students' level of English proficiency. To reach this important goal, the following recommendations are provided (S. Ferrara, personal communication, September 2007):

- States' ELP assessments are ongoing operational assessment programs, just like grade-level content area assessment programs. States should manage their ELP assessments accordingly.
- They should implement field testing procedures and replenish their item banks and operational test forms on a regular basis so that they do not over-expose current items, tasks, and test forms.
- They should conduct ongoing reviews of the alignment of items and assessment tasks with ELP standards and the psychometric characteristics of the item banks and test forms.
- States should plan and implement validity studies on an ongoing basis to examine current issues in assessing English language learners and ELP assessments that were discussed in this report.

Our main objective in this report has been to present information on the status of assessment of English language proficiency in the nation. We acknowledge our limitations in both content and scope of this overview. This

RETURN to TABLE of CONTENTS

is an extremely complex area in assessment and deserves more attention. We hope this presentation opens dialog on the importance and quality of ELP assessments nationwide. There is a great deal to learn from examining past experiences in this field as we consider the needs of the future. We welcome any comments and suggestions regarding the content of our work. We are hoping to incorporate suggestions and recommendations from readers into our future revisions.

We would like to take this opportunity to thank those who contributed to this work, including state departments of education, test publishers, scholars, and consortia members. They worked hard to assist us in providing information that is comprehensive and accurate.

References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Research, 33*(1), 4-14.

Abedi, J. (2006). Psychometric Issues in the ELL Assessment and Special Education Eligibility. *Teacher's College Record, 108*(11), 2282-2303.

Artiles, A. J., Rueda, R., Salazar, J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 7*(1), 283–300.

August, D., Francis, D., Hsu, H., & Snow, C. (2006). Assessing reading comprehension in bilinguals. *Elementary School Journal,* 107 (2), 221-239.

Bauman, J., Boals, T., Cranley, E., Gottlieb, M., and Kenyon D. (2007). Assessing Comprehension and Communication in English State to State for English Language Learners (ACCESS for ELLs®). In J.Abedi (Ed.), *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 81–91). Davis: University of California.

Bunch. M. B. (2006). *Final Report on ELDA Standard Setting*. Durham: Measurement Incorporated.

CDE, (1999). *English language development standards*. Sacramento, CA: California Department of Education.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles: National Dissemination and Assessment Center.

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

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

Garcia, E. E. (2005). *Teaching and learning in two languages: Bilingualism and schooling in the United States*. New York: Teachers College Press.

George, C., Linquanti, R. & Mayer, J. (2004). *Using California's English language development test to implement title III: Challenges faced, lessons learned.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Gottlieb, M., & Boals, T. (2006). *Using ACCESS for ELLs data at the state level: Considerations in reconfiguring cohorts, resetting annual measurable achievement objectives (AMAOs), and redefining exit criteria for language support programs serving English language learners* (WIDA Technical Report #3). Madison: WIDA Consortium, Wisconsin Center for Educational Research, University of Wisconsin.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: Johns Hopkins University Press.

Impara, J. C. and Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 355-368.

Jaeger, R. M. (1989). Certification of student competence. In: *Educational Measurement*, (Third Edition), Ed. by R. L. Linn, Washington, DC: American Council on Education, 485-511.

Kiplinger, V. L. (1996). *Investigation of factors affecting mathematics achievement in the eighth grade: Academic performance in Hawaii's public schools*. Unpublished Dissertation, University of Colorado at Boulder.

RETURN to TABLE of CONTENTS

Koffler, S.L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Education Measurement, 17*, 167-178.

Lara, J., Ferrara, S., Calliope, M., Sewell, D., Winter, P., Kopriva, R., et al. (2007). The English Language Development Assessment (ELDA). In J.Abedi (Ed.), *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 47–60). Davis: University of California.

Linquanti, R. (2004, April). *Assessing English-language proficiency under Title III: Policy issues and options*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Loomis, S. C. (2001). *Judging evidence of the validity of the national assessment of educational progress achievement levels*. Paper presented at the Annual Meeting of the Council of Chief State Schools Officers: Houston, TX.

Mathews, G., (2007). Developing the Mountain West Assessment. In J.Abedi (Ed.), *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 33–45). Davis: University of California.

Musick, M. D. (2000). *Can we talk?…About how to make education standards high enough in your state*. Atlanta, GA: Southern Regional Educational Board. Retrieved May 3, 2004 from http://www.sreb.org/main/highschools/accountability/settingstandardshigh.asp

NCELA (2007). *Glossary of terms related to linguistically and culturally diverse students*. Retrieved on May 8, 2007, from National Clearinghouse for English Language Acquisition: http://www.ncela.gwu.edu/

No Child Left Behind Act (2001). *Public Law 107-110. Title III—Language instruction for limted English proficient students*. Retrieved May 6, 2007 from http://www.ed.gov.

Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students. (2003, February). Draft of *Part II: Final Non-Regulatory Guidance on the Title III State Formula Grant Program—Standards, Assessments and Accountability*. U.S. Department of Education.

Rebarber, T., Rybinski, P., Hauck, M., Scarcella, R., Buteux, A., Wang, J., et al., (2007). Designing the Comprehensive English Language Learner Assessment (CELLA) for the Benefit of Users. In J.Abedi (Ed.), *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 63–79). Davis: University of California.

Robeinowitz, S. & Soto, E. (2006). *The technical adequacy of assessments for alternate student populations.* San Francisco: WestEd.

Sawaki, Stricker, & Oranje (2007). *Factor structure of an ESL test with tasks that integrate modalities*. Paper presented at the Annual Meeting of Educational Research Association. New Jersey: Educational Testing Service.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447-469.

Shepard, L., Camilli, G., & Williams, D. M. (1984) Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*(2), 93-128.

Teachers of English to Speakers of Other Languages, Inc. (1997). *ESL Standards for Pre-K-12 students*. Alexandria, VA: TESOL.

Teachers of English to Speakers of Other Languages, Inc. (2006). *PreK-12 English language proficiency standards*. Alexandria, VA: TESOL.

Texas Education Agency (2002). *State Accountability Data Tables, Base Indicators*; Texas Education Agency, *Secondary School Completion and Dropouts in Texas Public Schools 2001-02*, 2005, pg. 11.

Wiley, T. G. & Hartung-Cole (1998). Model standards for English language development: National trends and a local response. *Education, 119*(2), pp. 205-221.

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

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

RETURN to TABLE of CONTENTS

Appendix A

Overview of Existing English Language Proficiency Tests: Full Version of Chapter 7

Susan G. Porter and Jisel Vega

Over the years, many formal and informal assessments have been developed and used for the purpose of measuring English language proficiency of students whose home language is not English. Many of these assessments, however, do not meet the requirements specified in Title III of the NCLB Act. The law requires that these assessments:

- Measure annual progress in English language proficiency in the four domains of reading, writing, listening, speaking, and a separate measure for comprehension
- Be aligned to state-developed English language proficiency (ELP)/English language development (ELD) standards
- Relate to the state-adopted content standards, including those for reading/language arts and mathematics.

In addition, each state is encouraged to align its ELD standards to the state content standards and assessments.

In order to support the development of compliant English language proficiency tests, the U.S. Department of Education provided funding through the Enhanced Assessment Grant under NCLB, § 6112b. Four consortia received research support atnd four tests were developed: American Institute for Research (AIR) and the State Collaborative on Assessment and Student Standards (LEP-SCASS) collaborated on the English Language Development Assessment (ELDA); AccountabilityWorks in collaboration with several states developed the Comprehensive English Language Learner Assessment (CELLA); A World-Class Instructional Design and Assessment (WIDA) Consortium in collaboration with several states developed Assessing Comprehension and Communication in English State to State for English Language Learners® (ACCESS for ELLs); and Mountain West Assessment Consortium (MWAC) in collaboration with several states and Measured Progress developed the Mountain West Assessment (MWA).

DISCLAIMER: *The information presented in this chapter is as accurate and current as possible at time of publication. We welcome changes to the technical data so that it may be incorporated into the web-based version of this report over time. For the web-based version of this report, please see the UC Davis School of Education web site at http://education.ucdavis.edu/research/ELP_Assessment.html.*

RETURN to TABLE of CONTENTS

While many states are using an assessment developed by one of the consortia, other states developed their own assessments with the assistance of outside test developers. In several instances, states opted to use commercially available assessments, which are either "off the shelf" or augmented and aligned versions of assessments. Table 1 lists the assessments that states are currently using to meet Title III requirements. The table also indicates the name of the test developer(s) and the date of implementation.

The purpose of this chapter is to provide summary information on language proficiency assessments that states are currently using for Title III purposes. The development history and the technical aspects of *each* test will be briefly discussed in Appendix A.

Methodology

From August 2006 to April 2007 the research team used several methods to gather the information on each English language proficiency test. Initially, we searched state educational department and test developer/publisher websites and reviewed all documents pertinent to English language proficiency assessments for Title III purposes. Next, team members contacted Title III directors from all 50 states and the District of Columbia via e-mail. Title III directors were asked to share technical manuals, test administration manuals, alignment studies, website resources, and other relevant documents which would provide information on their state-adopted tests. Where necessary, team members also requested information from test developers/publishers and state assessment divisions. Follow-up phone calls were made to clarify test information, or to contact state representatives or publishers when prior attempts had not been successful. Informal email and phone conversations with test publishing companies/developers and state departments of education personnel served as additional sources of information in the data collection process.

The information provided in this chapter is as accurate and as current as possible at the time of publication. However, the following pages reflect only a "moment in time" snapshot of a dynamic process that is occurring nationwide as states and consortia continue their efforts to fully comply with English language proficiency assessment provisions within No Child Left Behind. In some cases, we had to rely upon unpublished and informal sources of data regarding assessment validity and reliability, standard setting, and item analysis. In many cases, states and test developers were still analyzing test data and/or technical manuals had not yet been published. For all of these reasons, changes to the technical data in this chapter are inevitable. Updates from the test developers and from state representatives can be incorporated in the web-based version of this report over time.[1]

Descriptions of Tests Currently Used by States for Title III Purposes

Summary information is provided for each assessment by test name, followed by grades covered, domains tested, publication date, and states using the assessment for Title III purposes. These summaries also include a description of the test purpose, score reporting information (i.e., proficiency levels), and a brief description of the test development. Where available, information about alignment activities and studies conducted during and after test development is included. Lastly, a discussion of the technical aspects of the test is included. Where available, information on item analysis, test reliability, validity, and freedom from bias is provided. The focus of the section on test technical properties is the types of psychometric tests conducted for each assessment; detailed results of each psychometric test are not provided. For more information on results of psychometric analysis for each assessment, the reader is referred to the test technical manual (as available).

As was indicated above, summary information for each individual assessment is provided in Appendix A. However, in this chapter, we present a summary that is characteristic of these assessments listed in Table 1 and discussed in Appendix A. Data from Table 1 will help the readers of this report gain a general idea of what ELP assessments are used by which states.

[1] See the UC Davis School of Education web site at http://education.ucdavis.edu/research/ELP_Assessment.html

RETURN to TABLE of CONTENTS

Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)

Grade Cluster(s): K; 1–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Dates): Alabama, Maine, and Vermont (2005). Delaware, Georgia, Illinois, New Hampshire, New Jersey, Oklahoma, Rhode Island, Washington D. C., and Wisconsin (2006). Kentucky, North Dakota, and Pennsylvania will implement ACCESS in 2007.

Test Purpose

Besides Title III accountability, ACCESS is also used:

- to determine student English language proficiency level to identify students who may qualify for English as a second language (ESL) or bilingual services
- to measure the annual progress of English language proficiency
- to evaluate the effectiveness of ESL/Bilingual programs or to enhance instructional programs for English language learners

Score Reporting

Scaled scores are provided in reading, writing, listening, and speaking. A comprehension score is a composite score based on performance in listening and reading. An overall composite score is based on scale scores across all four domains. Scale scores across the four domains are weighted differently: reading (35%), writing (35%), listening (15%), and speaking (15%).

The ACCESS overall composite score results in five proficiency levels that are denoted in a range of scores from 1.0 to 6.0. Scores from 1.0 to 5.9 are represented by the five proficiency levels entering, beginning, developing, expanding, and bridging. A sixth performance level descriptor denoted by 6.0, shows that the student has a level of conversational and academic language proficiency that is necessary for academic achievement at his/her grade level. ACCESS for ELLs® is a vertically scaled and horizontally equated assessment.

Test Development Summary

The Center for Applied Linguistics (CAL), under contract with the World-Class Instructional Design and Assessment Consortium (WIDA), began the development of ACCESS in fall 2003. Preliminary items were developed by teachers representing most of the states in the consortium in 2004. Items were piloted in two rounds in 2004. After the pilot tests, items were refined and two forms of the assessment were field tested in 2004 and 2005.

In April 2005, teachers and administrators determined cut scores for proficiency levels during the standard-setting process. For standard setting in listening and reading, they used a bookmarking procedure, while for writing and speaking they employed a modified body of work method. Specific information on scoring and standard setting is available in the technical manual.

Alignment to State Standards

The WIDA Consortium Steering Committee members, under the direction of Margo Gottlieb, developed the WIDA English language proficiency (ELP) standards with the first eight states in the WIDA consortium. As new states join the consortium, test developers continue to conduct studies to determine alignment between the WIDA ELP standards and academic standards of states newly entering the consortium. For further information, see the *English Language Proficiency Assessment and Accountability under NCLB Title III: A National Perspective* chapter in this report.

Technical Properties of the Test

Item analysis. Item analysis using Rasch methods was utilized as an empirical measure of item performance. Average item difficulty (average *p*-values) and both the average infit and outfit statistics were calculated to determine item performance. Specific information on item-level analyses is outlined in the first technical report (Kenyon, 2006).

Test Reliability. The reliability of the ACCESS test was examined using classical test theory (CTT), item response theory (IRT), and generalizability theory.

- CTT measurements of reliability conducted include calculation of Cronbach's alpha and the standard error of measurement (*SEM*).
- Reliability of the listening and reading portions of the test is shown by a reliability index, based on the concurrent calibration of all items using IRT, specifically Rasch techniques.

RETURN to TABLE of CONTENTS

- Inter-rater reliability for the reading section was conducted using three rating features: linguistic complexity, vocabulary use, and language control. A person and rater generalizability study was also conducted to determine rater reliability for the reading section. *G*-coefficients were also used to determine inter-rater reliability and generalizibility.
- Inter-rater reliability coefficient for the speaking section was determined, using Rasch methods. Specific information on reliability analyses conducted is available in the technical manual (Kenyon, 2006).

Test validity. Validity of this assessment was examined in several ways, including:

- Expert review: Content validity experts aligned ACCESS items to the WIDA ELP standards.
- Rasch measurement model fit and scaling: Misfitting items established in reading and listening sections following field testing.
- Concurrent validity determined: Pearson correlation coefficient was calculated by comparing ACCESS for ELLs® scores with four other test scores: the Language Assessment Scale (LAS), The Language Proficiency Test Series (LPTS), the IDEA Proficiency Test (IPT), and The Maculaitus II (MAC II). For a fuller description, refer to the ACCESS for ELLs® technical manual.
- Correlations between scale scores across item domains calculated. Specific information on test validity is outlined in the ACCESS for ELLs® technical manual.

Differential item functioning (DIF) was conducted by gender and ethnicity (Latinos were the focal group while all others consisted of the reference group). For dichotomously scored items, the Mantel-Haenszel chi-square statistic with the Mantel-Haenszel common odds ratio that is converted to the Mantel-Haenszel delta scale, was used, following Educational Testing Service (ETS) guidelines.

Polytomously scored items were measured through the Mantel chi-square statistic and the standardized mean difference procedures, following ETS guidelines.

For specific information on analyses conducted to ensure freedom from bias, consult the ACCESS for ELLs® technical report.

Technical Reports

Kenyon, D., MacGregor, D., Jeong, Ryu, J., Cho, B., & Louguit, M. (September 2006). *Annual technical report for ACCESS for ELLs®, series 100, 2005 administration* Annual (Tech Report No. 1) Abridged version for TAC discussion (complete but with output for grades 3–5 only) (DRAFT). Center for Applied Linguistics.

Kenyon, D. (August 2006). *Development and field test of ACCESS for ELLs®* (Technical Report No. 1) Center for Applied Linguistics.

Gottlieb, M. & Kenyon, D (August 2006). *The bridge study between tests of English language proficiency and ACCESS for ELLs® part I: background and overview* (Technical Report #2.) Center for Applied Linguistics.

Gottlieb, M. & Kenyon, D (August 2006). *The bridge study between tests of English language proficiency and ACCESS for ELLs® part II A: IPT results* (Technical Report #2.) Center for Applied Linguistics.

Gottlieb, M. & Kenyon, D (August 2006). *The bridge study between tests of English language proficiency and ACCESS for ELLs® part II B: LAS results* (Technical Report #2.) Center for Applied Linguistics.

Gottlieb, M. & Kenyon, D (August 2006). *The bridge study between tests of English language proficiency and ACCESS for ELLs® part II C: LPTS results* (Technical Report #2.) Center for Applied Linguistics.

Gottlieb, M. & Kenyon, D (August 2006). *The bridge study between tests of English language proficiency and ACCESS for ELLs® part II D: MAC II results* (Technical Report #2.) Center for Applied Linguistics.

Arizona English Language Learner Assessment (AZELLA)

Grade Cluster(s): K; 1–2; 3–5; 6–8; 9–12
Domains Tested: Grades 1-12: reading, writing, listening, speaking, and writing conventions. Pre-literacy level (kindergarten): pre-reading, prewriting, and speaking.
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (State Adoption Dates): Arizona (Fall 2006)

RETURN to TABLE of CONTENTS

Test Purpose

AZELLA is used to determine student placement in ELL programs, student progress, and ELL reclassification to fluent English proficient (FEP) resulting in exit from the ELL program.

Score Reporting

Grade clusters are labeled in the following manner for test reporting purposes: Pre-literacy – kindergarten, Primary – grades 1–2, Elementary – grades 3–5, Middle grades – grades 6–8, and High School – grades 9–12. The prewriting and speaking domains for the Pre-literacy (kindergarten) level and the speaking domain for grades 1–12 are scored on site by the ELL site coordinator/tester. Students are assigned one of five proficiency levels based on each domain score and student's composite score: Pre-Emergent, Emergent, Basic, Intermediate, and Proficient.

Test Development Summary

The AZELLA is an augmented and aligned version of the Stanford English Language Proficiency (SELP) assessment developed by Harcourt Assessment Inc. New items were field tested in October 2005. Items that performed well on this field test were retained and a forms field test was conducted during February and March 2006. The resulting form, AZ-1, has been used since August 2006 to test English language learners in Arizona. During 2006–2007 an alternate form of the AZELLA, AZ-2, will be developed. This form will be an augmented version of SELP Form C for each grade level cluster. A Kindergarten Form B will also be published. Form AZ-2, will be fully operational by August 2008.

Harcourt Assessment facilitated standard-setting meetings for AZELLA Form AZ-1 in Phoenix in June 2006. Information from the AZELLA item field test, AZELLA forms field test, Arizona ELL practitioners, and the SELP were examined in order to determine the language proficiency levels and cut scores. The AZELLA is a vertically scaled assessment.

Approximately 125 new or modified test items are used on the AZELLA Form AZ-1. However, the AZELLA retains at least 30% of the SELP items in each domain from Form A in order to maintain the SELP vertical scale.

Alignment to State Standards

In 2006, Aha! Inc. conducted an alignment study between the AZELLA, Form AZ-1 and the Arizona English Language Learner Proficiency Standards. This study concluded that the alignment was high; however, gaps and under-assessed areas were found between the standards and the AZELLA. These results are being used to assist in the development of the test blueprint for Form AZ-2 to ensure adequate coverage of content.

Technical Properties of the Test

Item analysis. Information on item analysis was not available.

Test reliability. Several indices were used to ascertain the reliability of the AZELLA including both classical test theory (CTT) and item response theory (IRT). CTT approaches include the internal consistency (Cronbach's coefficient alpha) and the standard error of measurement (*SEM*). The reliability of the total composite for grades K and 1 is 0.93, and the reliability for grades 3 through 12 is in the high 90s. In addition, the conditional *SEM* based on IRT was calculated.

Test validity. Information on test validity and analyses conducted to ensure from bias was not available.

Technical Reports

AZELLA technical details, including test design and development, item-level statistics, reliability, validity, calibration, equating, and scaling, will be available in 2007 when the 2006 AZELLA technical report is released by Harcourt Assessment.

Seibert, M., Turner, C., & Pimentel, S. (June, 2006). R*eview of AZELLA form AZ-1 alignment to Arizona English language learner proficiency standards report.*

Phoenix, AZ: Arizona Department of Education. Unpublished draft technical report 2006, Harcourt Assessment, Inc.

Colorado English Language Assessment (CELA)

Grade Cluster(s): K–1; 2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Adoption Dates): Colorado (spring 2006)

RETURN to TABLE of CONTENTS

Test Purpose

The CELA is used to measure annual progress of English language proficiency.

Score Reporting

Scale scores are provided in reading, writing, listening, and speaking, and an overall score is calculated. In addition, a comprehension score is derived from parts of listening and reading results. An oral language score is derived from listening and speaking domains. Scale scores are used to designate five levels of proficiency, ranging from the lowest proficiency of 1 to the highest proficiency of 5.

Test Development Summary

The Colorado English Language Assessment is the Language Assessment System Links (LAS Links) Form A. Colorado began to use this assessment in 2006. In 2007, some minor changes were made to the assessment, including changes to procedures, book covers and the biographical page; however, no changes to the test content were made. Beginning in 2008, new items will be introduced in the CELA to enhance the alignment to the Colorado ELD standards. The percentage of newly developed items is expected to increase by a quarter of the test through 2010. For additional information on test development and standard setting, please see the summary on LAS Links in this chapter or in the CELA technical manual.

Alignment to State Standards

According to the Colorado Department of Education, the CELA is 80–85% aligned to Colorado's English language development standards. CTB/McGraw Hill and the Colorado Department of Education are working together to increase the number of aligned test items from 2008 through 2010. An alignment study was not available.

Technical Properties of the Test

Item analysis; test reliability; test validity. Please see the summary on LAS Links in this chapter or the CELA technical manual for test validity and freedom from bias information.

Technical Report

CTB/McGraw-Hill (2006). *LAS links technical manual.* Monterey, CA: CTB/McGraw-Hill LL.

California English Language Development Test (CELDT)

Grade Cluster(s): K–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2001
State(s) Using This Test for Title III Accountability (Implementation Date): California (2001)

Test Purpose

In addition to Title III accountability, CELDT is used to: identify students as English language learners in grades K–12, determine students' level of English language proficiency, and assess students' annual progress in acquiring English across domains.

Score Reporting

Scaled scores are provided in reading, writing, listening, speaking, and listening/speaking. A comprehension score is derived from combining the listening and reading scale scores. An overall score is calculated from a combination of all domains. There are five proficiency levels: Beginning, Early Intermediate, Intermediate, Early Advanced, and Advanced.

Test Development Summary

The CELDT was developed by CTB/McGraw Hill (CTB) under contract with the California State Department of Education (CDE). In 1997, legislation authorized the CDE to develop ELD standards and a language proficiency assessment that would be used statewide. In 1999, California ELD standards were adopted. The first version of the CELDT assessment consisted primarily of items from the Language Assessment Scales (LAS) tests with some new items developed specifically for California by CTB. This test version was field tested in fall 2000. Data from the field test were used to select items and create the operational forms of the test. The first bookmark standard-setting study was conducted in spring of 2001 to determine cut scores that would define the five proficiency levels. The first test administration took place between May and October 2001. The CELDT has been updated yearly since 2001. Subsequent versions have gradually replaced LAS items with test items that are aligned with the California ELD standards. Table 1 matches CELDT test forms and their administration dates.

RETURN to TABLE of CONTENTS

Table 1. CELDT test forms and their administration dates

Form	Administration Dates
Form A	2000–2002
Form B	2002–2003
Form C	2003–2004
Form D	2004–2005
Form E	2005–2006
Form F	2006–2007
Form G	2007- 2008

Note: During 2008–2009 the letter form identifier will no longer be used.

In February 2006, CTB/McGraw-Hill conducted a second standard setting study with education experts from California (classroom teachers, content specialists, school administrators, and others designated by the CDE). The bookmark standard-setting procedure (BSSP) was used to set new performance-level cut scores on the CELDT. These—and a common scale for the CELDT—were implemented in July 2006. The CELDT uses a common scale in order to facilitate tracking of growth across adjacent grade spans and proficiency levels.

Alignment to State Standards

A two-part alignment study was conducted by CTB/McGraw-Hill in conjunction with the California Department of Education (Murphy, Bailey, and Butler, 2006) to determine the linkage and/or degree of alignment between: the state-adopted ELD standards *and* the CELDT test; the language demands of the state-adopted ELD standards *and* the state-adopted content standards; and the California ELD Standards *and* the state-adopted content area assessments. The methodology for these studies used expert raters who conducted document reviews of the standards and frameworks and compared them with the blueprint and specific test items. Degree of alignment between the standards/frameworks and Form E of the CELDT test were based upon the following dimensions:

1. ratability,
2. domain,
3. complexity, and
4. language demands.

For the last dimension, *language demands*, Murphy et al. used academic language frameworks developed by Stevens, Butler, and Castellon-Wellington (2000) and Scarcella & Zimmerman (1998) to identify the following three categories of words needed for content area knowledge: (a) high-frequency general words, or words used regularly in everyday or social contexts; (b) non-specialized academic words, or words that are used in academic settings across content areas; and (c) specialized content-area words, or academic words unique to specific content areas (i.e., Math and Social Science).

Percentage of alignment was computed in these four dimensions for each grade level. Using overall percentage of alignment by test and grade span, a frequency distribution was developed to determine the relative strength of alignment between the ELD Standards and the CELDT, between the CELDT Standards and the state-adopted content area tests.

For the linkage (at the objective level) between the state content standards and the ELD Standards, the overall *ratability* was determined to be 74%. The linkage between the content and ELD standards for the *domain* dimension revealed similar percentages for listening. On the other hand, roughly two-thirds of ELD standards required speaking and writing while only one third of the content standards required these domains. Reading ranged from 23.0% to 34.4% in the ELD standards, compared to 15.3% to 42.9% in the content standards.

Across the grades, the ELD standards were coded at lower *language complexity* levels than the content standards. For example, in Grade 2, over half of the ELD standards were considered *low* complexity while about one third of the content standards were considered *low* complexity. Almost 4% of the Grade 2 content standards were rated as *high* compared to less than 1% of the ELD standards. Crosswalk analyses showed that Grades 2, 7, and 9 showed the strongest linkages across all language demands at most levels of complexity. Grade 5 showed particular weakness in linguistic skills for the content areas. Overall alignment in the area of language demands ranged from 3% to 35% alignment between the CELDT and the content area assessment by grade CELDT domains.

Technical Properties of the Test

Item analysis. The most recent item analysis was conducted on Form C of the CELDT. *P*-values were calculated as a measure of item difficulty. Point-biserial correlations were calculated as a measure of item discrimination. Differences between *p*-values for the annual administration data and the initial identification data were calculated, and correlations between multiple choice and constructed response items were determined. Specific item-level statistics for each analysis are provided in the technical report manual.

Test reliability. Test-retest reliability studies conducted for Form C of the CELDT showed that test-retest reliability was determined to be between .85 and .90. The test

RETURN to TABLE of CONTENTS

developer notes that this coefficient was derived from estimates yielded from field testing of items embedded within test versions, not through the usual means of administering multiple parallel forms of a test to the same student. In addition, the standard error or measurement is provided as an indication of test reliability. Standard errors range from 17 to 26 points across all grades and subject areas in scale score units. Rater consistency and reliability was also examined for the reading portion of the assessment. Additional information on reliability is located in the technical manual.

Test validity. Criterion-related validity was assessed in an independent study conducted by Katz (2004). This study found that there were only moderate correlations between the English learner CELDT scores and their scores on the state-adopted content area assessment in reading (the Stanford 9). Overall correlation on student scores in reading for these two assessments ranged from .71 for second graders to .44 for tenth graders. There was an overall trend for the correlation coefficient to decrease with the higher grade levels.

Technical Reports

CTB McGraw Hill (2002). *California English language development test: Technical report for the California English language development test (CELDT), 2000–2001 (Form A)*. Retrieved March 29, 2007, from http://www.cde.ca.gov/ta/tg/el/resources.asp

CTB McGraw Hill (2003). *California English language development test: Technical report for the California English language development test (CELDT), 2002–2003* (Form B). Retrieved March 29, 2007, from http://www.cde.ca.gov/ta/tg/el/resources.asp

CTB McGraw Hill (2004). *California English language development test: Technical Report for the California English Language Development Test (CELDT), 2003–2004* (Form C). Retrieved March 29, 2007, from http://www.cde.ca.gov/ta/tg/el/resources.asp

Comprehensive English Language Learning Assessment (CELLA)

Grade Cluster(s): K–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Date): Florida (fall 2006) and Tennessee (spring 2005).

Test Purpose

The CELLA was created to meet accountability requirements outlined in Title III for English language learners and to:

- Measure student progress over time. Proficiency levels of individual students may used to make placement and exit decisions for English as a second language (ESL) and bilingual education programs.
- The CELLA also provides information on individual student strengths and weaknesses in English language proficiency, which may be used for diagnostic purposes.

Score Reporting

CELLA provides scores in listening/speaking, reading, and writing as well as a comprehension score. A total score, which is derived from the listening/speaking, reading, and writing scores, is provided. Anchor points are provided as a general indicator of student proficiency. In listening and speaking scale scores are matched to four anchor points. Anchor point 1 demonstrates the lowest level of proficiency; anchor point 4 demonstrates the highest level of proficiency. In reading and writing, scale scores are matched to five anchor points, with anchor point 1 demonstrating the lowest level of proficiency and anchor point five demonstrating the highest level of proficiency. It was recommended that individual states conduct their own standard-setting studies to establish proficiency levels and cut scores (see *Test Development Summary* for information regarding standard-setting studies by individual states).

Test Development Summary

The CELLA was developed by Educational Testing Service (ETS), Accountability Works, and a consortium of five states: Florida, Maryland, Michigan, Pennsylvania, and Tennessee. Items were developed and field tested between October 25 and November 8, 2004, although testing was extended in some schools to accommodate requests for later testing dates. Field testing was conducted in the states of Florida, Maryland, Pennsylvania, and Tennessee with students in grades K–12. Three field test forms were administered at each grade cluster and items most appropriate were selected to create the final forms.

RETURN to TABLE of CONTENTS

Scale anchoring was conducted by ETS in order to determine exemplar items that best demonstrate how students perform at different points on the vertical scale. These exemplar items were transformed into behavioral descriptions by content experts. Individual states then used these descriptors in their own standard-setting studies.

Standard setting was conducted in the State of Florida in winter 2006. An Educator Panel Workshop was conducted to develop proficiency-level descriptors (PLDs) to describe what is expected at each level of language proficiency. Panel members used PLDs to define benchmarks used in the standard-setting process. Using a bookmarking procedure, three recommended cut scores at each grade level cluster for Oral skills, reading and writing were developed resulting in four levels of student proficiency: beginning, intermediate, advanced and proficient.

In 2006, ETS directed a standard-setting study in Tennessee with ESL educators from across the state. The bookmarking process was used to recommend cut scores for each form of the test at one grade level. ETS conducted statistical analysis and presented impact data as well as estimate cut scores for other grades and grade level spans. The resulting proficiency levels were Beginner, High Beginner, Intermediate, High Intermediate, and Advanced. Additional information on scoring and standard setting is outlined in the technical report.

Two complete operational forms, Form A and B, for each grade cluster were published in 2005. Tennessee began administering the CELLA in spring 2006, while Florida began using the assessment in fall 2006.

Alignment to State Standards

Test items in the CELLA are aligned to the CELLA proficiency benchmarks. During test development, Accountability Works (AW) assisted content experts in alignment analyses. These analyses found high levels of alignment between the CELLA benchmarks and ESOL standards for Florida, Michigan, and Pennsylvania. A similar alignment study was performed between consortium state academic Reading/Language Arts standards to the CELLA benchmarks. An alignment study between the State of Tennessee's English as a second language (ESL) standards and CELLA proficiency benchmarks could not be located.

Technical Properties of the Test

Item analysis. *P*-values were calculated to determine item difficulty of multiple choice items. For constructed response items, an equivalent index consisting of the mean item score divided by the maximum possible item score was used to determine item difficulty. To assess item discrimination, the correlation between students' item scores and their total test scores was used for both item types.

On average, the *p*-values for items in the listening section were in the low 0.70s. The *p*-values for the speaking items were in the low 0.70s on average. *P*-values showed that the overall reading items were more difficult than the listening and speaking items. In Level A, the reading items were shown to be more difficult than items in the other domains. Items were also analyzed in an item response theory (IRT); the three-parameter logistic model was used for multiple-choice items and the generalized partial-credit model was used for the constructed-response items. Additional information on item analyses is provided in the technical report.

Test reliability. Reliability was estimated using Cronbach's alpha for listening and reading sections; a stratified coefficient alpha was derived from field test scores in the reading and speaking sections. Reliability was reported by content area (reading, etc.) and by form (Form A, B2, etc.). Internal consistency ranged from a low of .76 to a high of .95. The standard error of measurement was reported in addition to the IRT based approach, the conditional standard error of measurement (*CSEM*). Additional information on test reliability is outlined in the technical report.

Test validity. Validity was ensured through states' review and approval of the proficiency standards, test blueprints, item specifications, and items used in the tests. In addition, expert review of the assessment was conducted to ensure that all items match the proficiency benchmarks and specification and that all forms of the assessment match test blueprints. Additional information on test validity is outlined in the technical report.

CELLA items were reviewed by trained ETS reviewers for fairness and sensitivity. In addition, Differential Item Function (DIF) was carried out by gender using the Mantel-Haenszel (Mantel & Haenszel, 1959) and the Standardization (Dorans & Holland, 1993) approaches. Additional information on analyses conducted to ensure freedom from bias is outlined in the technical report.

RETURN to TABLE of CONTENTS

Technical Reports

Comprehensive English language learning assessment technical summary report educational testing service (2005). Princeton, New Jersey: Educational Testing Service.

Dakota English Language Proficiency Assessment (Dakota ELP)

Grade Cluster(s): K–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Date): South Dakota (2006)

Test Purpose

This assessment is used to:

- Annual review of progress of ELLs for state and Title III accountability purposes.
- As a diagnostic tool to assist teachers in determining which instructional standards they must focus on so that ELLs fully acquire English language proficiency.

Score Reporting

See the summary on the Stanford English Language Proficiency (SELP) in this chapter for more information on the scoring of this assessment.

Test Development Summary

The Dakota ELP is an augmented and aligned version of the SELP Assessment Form A developed by Harcourt Assessment Inc. Items on the Dakota ELP were developed following an alignment study conducted in February 2005. The first field test of the Dakota ELP was conducted in September 2005. Items which performed well were chosen to be included on the final forms of the assessment. The test was fully administered in the State of South Dakota for the first time in February 2006.

Alignment to State Standards

In spring 2003, South Dakota began to use Form A of the SELP to comply with Title III and South Dakota state requirements for testing. South Dakota adopted English language development standards in 2004, and subsequently, an alignment study was conducted by H. Gary Cook (Cook, 2005) to determine alignment of these standards and the state content standards to the SELP. The Webb alignment methodology was used for this study. The findings of this study showed weak alignment between South Dakota's content standards and the listening and writing domains for grades K–2. In addition, the SELP had limited alignment to the state's math standards in grades K–2. The study also showed that alignment of the SELP with the state content standards was weak for grades 3–5 when compared to other grade clusters. In grades 6–8 there was satisfactory alignment of the SELP to ELD standards, with reading being the weakest area of alignment. Alignment between SELP items for grades 9–12 (?) and the ELD standards were satisfactory, with the speaking domain having the weakest alignment with the standards. It was determined that the SELP was not aligned sufficiently to South Dakota's English language development standards in mathematics so the test was augmented to address deficiencies in this area. These items were added to the SELP 2006 assessment.

The Buros Institute, in conjunction with the South Dakota Department of Education and Harcourt Assessment, led standard setting to determine performance levels in May 2006. Grade- and content-level teachers participated in the standard-setting process. Cut points for each proficiency level were established through item level analysis. Raw scores and corresponding scaled scores were then assigned to each proficiency level. The Dakota ELP is vertically scaled. Specific information on standard setting can be found in the technical manual.

Technical Properties of the Test

Item analysis. *P*-values were calculated to determine item difficulty and point-biserial correlations were calculated as a measure of discrimination. Test items had *p*-values between 10–90%, with items having a *p*-value close to 50% favored. Those items with correlations larger than .3 were considered for use. However, those items with point-biserial correlations close to zero, zero, or negative were not used. The item response theory (IRT) approach to item evaluation was also undertaken. The Rasch item-response model was estimated using the joint maximum likelihood (JML) method. Specific information on item analyses is located in the technical report.

Test reliability. Reliability was estimated using Cronbach's alpha coefficient and the classical test theory (CTT) standard error of measurement (*SEM*). The range of the reliabilities of the total test across grades K–12, as provided by Cronbach's alpha, ranges from 0.905 to 0.955.

In addition, Livingston and Lewis's method was used to obtain measures of the decision accuracy and consistency of the classifications of the five performance levels. The accuracy of the decision to classify the students into Proficient or above versus Intermediate or below for the total test (a composite score of all domains) ranged from 76.9% to 99.7% across all grades. The consistency of the decision ranged from 72.2% to 99.7%. In all cases, decision accuracy was greater than decision consistency. Specific information on reliability is located in the technical report.

Test validity. The technical report focused on content and constructs validity. To ensure the content of curricular validity of the *Dakota ELP*, an alignment study was conducted to verify that the assessment is aligned with the state ELL standards for each corresponding subject and grade level. In addition, the correlations between the domains and the total test were calculated and ranged from .357 to .889. Most correlations between the domains and the total test were larger than .600. For specific information on test validity refer to Dakota ELP technical report.

Prior to field testing, items were reviewed by Harcourt assessment experts and experts chosen by the South Dakota Department of Education to ensure freedom from bias of items. Specific information on freedom from bias is available in the technical report.

Technical Report(s)

South Dakota Department of Education. (2006). *South Dakota state English language proficiency assessment Dakota ELP technical report: 2006 Spring administration* (Draft). San Antonio, TX: Harcourt Assessment, Inc.

English Language Development (ELDA) K–2 Assessment

Grade Cluster(s): K–2
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Arkansas (spring 2007), Iowa (spring 2006), Louisiana (2005, 2006), Nebraska (2005, 2006), Ohio (spring 2006; K–2 only), South Carolina (2005, 2006, 2007), Tennessee (2007), West Virginia (spring 2005)

Test Purpose

Developed to satisfy the requirements of Title III of the No Child Left Behind Act, ELDA K–2 is also used to determine English language proficiency levels for children from kindergarten through grade 2.

Score Reporting

ELDA K–2 consists of two inventories; one for kindergarten, and the other for first and second grade. Teachers record the scores for each item in the individual student's test booklet. For some states Measurement Incorporated provided raw and scale scores for both inventories yielding a proficiency level for each domain and for an overall language proficiency score. One of five language proficiency levels is determined for each domain, Comprehension and for an overall language proficiency level: 1-Pre-functional, 2-Beginning, 3-Intermediate, 4-Advanced, and 5-Fully English Proficient.

Test Development

Please see entry on the ELDA (grades 3–12) for information on primary test developers, consortium information, and initial ELD standards development. The consortia members—CCSSO, LEP-SCASS, AIR, Measurement Incorporated, and C-SAVE—determined English language proficiency assessments for younger English learners should rely on observational data in natural settings. For this reason, the consortia undertook the development of a separate test blueprint for testing English language proficiency for kindergarten through grade 2 students.

In November 2003, project members of the K–2 advisory sub-committee met with AIR staff to review the consortium states' ELP and content standards and select those appropriate to students in kindergarten through 2nd grade. The subcommittee, in consultation with experts in early childhood education, developed a final set of ELP and content standards appropriate for this grade range.

In February 2005, Measurement Incorporated coordinated the test item development. Classroom teachers developed constructed response items for the ELDA K–2. Each item was designed to be a statement regarding a specific student behavior. As part of this scope of work, anchor items from the ELDA for grades 3—5 were chosen for inclusion in the K–2 assessment, these items were included in the K–2 assessment so that scores from this assessment would link to those from the ELDA 3–12. The initial bank of items was reviewed for bias and content and then forwarded to CCSSO for final approval.

RETURN to TABLE of CONTENTS

Data from the fall, 2005 field test administration was analyzed to conduct item analysis and to determine preliminary cut scores for each of the five proficiency levels. During the 2005 field testing, each teacher who administered the K–2 inventory was asked to rate the proficiency of each student on a scale of 1–5. These ratings were used to determine initial cut scores for each of the five proficiency levels (see Score Reporting, above) by finding the mean raw scores of all students rated for a particular level, then finding the midpoint between the mean raw scores of students rated at adjacent proficiency levels.

Based upon feedback from consortia state representatives, it was determined that the ELDA K–2 was too long and difficult to administer for the age group for which it was intended. Measurement Incorporated undertook the task of shortening the instrument for the spring 2006 test administration. The amended inventories were reviewed in December 2005 and approved by the consortia in January 2006.

Measurement Incorporated conducted a follow-up standard setting in January 2006. Prior to this second standard-setting study, CCSSO updated the Performance Level Descriptors (PLDs) for the five proficiency levels of the K–2 inventory. These updated PLDs were referenced by an expert panel to examine student work and to classify each student as being in one of the five proficiency levels listed above.

Alignment to State Standards

Similar to the ELDA 3–12, the ELDA K–2 was based upon a set of English language development standards that were developed by the consortia. Please see the technical manual and department of education websites for alignment studies between the ELDA and participating states' adopted content standards and ELD/ELP standards.

Technical Properties of the Test

Item analysis. Because of the nature of the test, item analysis consisted of calculating the means and standard deviations of responses for each item (each student could earn a score of 0, 1, 2, or 3 per item), as well as the correlations between students' item scores and individual students' total scores. This item analysis provided information about item difficulty.

A second set of calculations determined the effect that each item had upon the total score by removing it from the average total score. This provided the test developers with information on item functioning and the impact that a single test item had on the total test.

Test reliability. Phi coefficients (equivalent to KR-21) and generalizability coefficients were derived from both the 2005 field test administration (long version) and the spring 2006 field administration (shortened version). Reliability data on the spring 2006 field test administration of the shortened ELDA K–2 Inventory showed that the generalizability coefficients (equivalent to coefficient alpha or KR-20) were between .93 for the Writing Inventory and .96 for the Reading Inventory, showing high internal consistency.

Test validity. The results of the teacher rating scores (see Test Development, above) were correlated with the 2005 field test results of the ELDA K–2. These data show that correlations between teacher ratings and scores on the Reading and Speaking Inventories (for all grades) were both .68, while the correlations between teacher ratings and listening were .57 and .58 for reading (grades K through 2).

Overall test validity was also monitored through expert judgment and teacher feedback of alignment between ELDA K–2 test items and performance level descriptors (PLDs). Comments from teachers who administered the field test forms also informed test developers of the relevance of the test items to classroom instruction.

Bias reviews were conducted throughout the item development and test review process. Additionally, individual item statistics were studied, following the 2005 field test administration, to determine whether certain subgroups responded significantly differently from other subgroups. These subgroup comparisons included gender, grade, and race.

Technical Report

Published by AIR (2005). Available on the Council of Chief State School Officers' website: http://www.ccsso.org/projects/ELDA/Research_Studies/

English Language Development Assessment (ELDA)

Grade Cluster(s): 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2005

RETURN to TABLE of CONTENTS

State(s) Using This Test for Title III Accountability (Implementation Date): Arkansas (2007), Iowa (spring 2006), Louisiana (spring 2005), Nebraska (spring 2005), South Carolina (spring 2005), Tennessee (spring 2007), West Virginia (spring 2005)

Test Purpose

The ELDA was developed to meet the English language proficiency assessment requirements outlined in Title III of No Child Left Behind and:

- to assess a construct of "academic English" in the domains of reading, writing, listening, and speaking.
- to measure progress in the development of English language proficiency across three grade clusters within grades three through twelve.

Score Reporting

Scale scores from the four domains are used to determine English Language proficiency by levels (Pre-Functional, Beginner, Intermediate, Advanced and Fully English Proficient) in all four domains (reading, writing, listening, and speaking). Additionally, a composite score in overall English proficiency is derived from domain scores in the four domains. A comprehension score is calculated from the listening and reading test scores.

Test Development Summary

The ELDA was developed as part of an Enhanced Assessment Grant under Title VI of NCLB issued to the State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS) consortium. This consortium was led by Nevada in collaboration with other members of the consortium: the Council of Chief State School Officers (CCSSO), the American Institute for Research (AIR), Measurement Incorporated (MI), and the Center for the Study of Assessment Validity and Evaluation (C-SAVE) at the University of Maryland.

AIR, in conjunction with an expert panel, developed a set of English language development standards prior to the development of the language proficiency test. This expert panel and AIR convened in December 2002 to compare English language development standards adopted by states in the consortium, as well as standards adopted by California, Nevada, New Jersey, and Texas. The expert panel adapted these state-adopted standards, keeping assessed grade levels and the needs of those states in the LEP-SCASS consortium in mind. Many states participating in the consortium later adopted these revised ELP standards, or used this set of standards as a basis upon which to develop their own English language proficiency standards. The expert panel and AIR determined the five proficiency levels (Pre-functional, Beginning, Intermediate, Advanced, and Fully English Proficient) and performance level descriptors for each language domain, Comprehension and composite in the process of developing the performance standards.

American Institutes for Research, in conjunction with CCSSO, developed benchmarks based upon the adopted standards for the reading, listening, and writing domains. (Standards for which benchmarks could not be developed were considered not acceptable or testable). Benchmarks then guided the development and review of individual test items that were included in the development of test forms (AIR, 2005). Item development occurred in February 2003 and was conducted by AIR using a pool of content experts and experienced item writers. AIR then conducted a three-step item review process which examined draft items for possible bias, clarity and grammar, and finally, for their content validity and match to the ELDA standards. These items were reviewed by consortia members to determine the items to be field tested.

From the selected test items, AIR created two field test forms (A and B) for each grade cluster (3–5, 6–8, and 9–12) and for each field domain (reading, writing, listening, and speaking). To vertically link the grade cluster test forms, selected test items were included across grade clusters. The test forms were field tested in March 2004 by Measurement Incorporated. Data analysis of the field test results by AIR (reported in the validity section, below) showed that the test items did not distinguish sufficiently between students at differing proficiency levels. This study led to the refinement of the recorded prompts provided for the speaking test, as well as changes in the scoring rubrics.

A second field test/operational test administration was conducted in 2005. Operational test data were gathered from five participating states: Iowa, Louisiana, Nebraska, Ohio, South Carolina, and West Virginia. Another five states (Georgia, Indiana, Kentucky, New Jersey, and Oklahoma) participated in the 2005 ELDA field test for grade clusters 3–5, 6–8 and 9–12.

A preliminary standard-setting activity was conducted in 2004 with the 2004 field test data to determine cut scores for the final field test administration. Measurement

RETURN to TABLE of CONTENTS

Incorporated conducted another standard-setting study in August 2005 to determine proficiency levels and cut scores for the ELDA. Expert panel members representing the participating states were divided into four committees to review student responses to the test items, based upon the 2005 field test administration.

Committees 1 through 3 examined responses to the items in the reading, writing, and listening domains by grade cluster. Committee 4 reviewed responses across all grades to the speaking domain items. Committee members used a bookmarking procedure to determine cut-scores for each of five proficiency levels for each grade cluster using the performance level descriptors (PLDs) determined for each domain. Committee members were also informed by the Rasch statistic for each test item, which was included in the difficulty-ordered test booklets provided for the standard-setting activity. A subgroup selected from Committees 1 through 4 formed the Articulation Committee; this committee reviewed all of the preliminary cut scores determined by the other committees to make a final determination regarding proficiency-level cut scores for all domains within each grade cluster. The committee members determined that reading and writing would be weighted in the computation of Comprehension and Composite levels. The results of the standard setting was presented to and approved by the consortia members.

West Virginia developed and adopted a revised version of the ELDA, the West Virginia Test for English Language Learning (WESTELL). Iowa recently made significant changes to the ELDA, particularly in the format and administration of this instrument. This assessment was adopted by West Virginia in 2007 and will is called the Iowa-ELDA (I-ELDA).

Alignment to State Standards

Following the development of the ELDA language proficiency standards, AIR developed *benchmarks*, or specific statements of what students should know and be able to do as measures of progress toward meeting a standard. Benchmarks were developed for all domains, with the exception of the speaking standards. The resulting standards and benchmarks were then used in the determination of test specifications and mapped onto each test item as it is developed. While participating states developed ELP standards that were based upon the consortia-adopted ELP standards, it is unclear whether individual states conducted their own alignment studies between their state-adopted ELP/ELD standards, the state-adopted content standards, and the ELDA assessment, with the exception of Nebraska.

In December 2004, the Nebraska Department of Education conducted a correlation study of their K-12 Guidelines for English Language Proficiency, the ELDA, and the state-adopted academic content standards. Teams of ELL practitioners and district administrators participated in the connections study. Documentation forms were then developed that allowed for the identification of the item, connection to standard, proficiency level, and review comments. The state standards were reviewed for continuity, and then compared with the ELDA items across grade levels and domains. The reviewers substantiated a correlation between the ELL standards, assessment tools and the state standards.

In Louisiana, ELL practitioners, content specialists and district administrators conducted a review of the state adopted English Language Development Content Standards and benchmarks based on the ELDA standards to determine the linkage between the English language development standards and state adopted academic content standards. This document has been used in professional development for ELL practitioners and content teachers.

Technical Properties of the Test

Item analysis. AIR staff conducted both classical test (CT) statistical analyses and item response theory (IRT) statistical analyses with data from the 2004 and 2005 field test administrations on the ELDA. These data analyses provided information on each item, as well as information on the validity and reliability of the overall testing forms. The CT analyses of the 2005 ELDA test items showed that Test difficulties range from p=0.54 for reading in grade cluster 6–8 to p=0.81 for speaking in grade clusters 3–5 and 9–12. Test difficulties are comparable across grade clusters in each skill domain. Adjusted biserial and polyserial correlations ranged between r=0.47 and r=0.87. The average omit rate was 3.11% across all skill domains, grade clusters and test forms. The highest number of items was omitted in speaking in grade cluster 6–8 (11.97%); the lowest number of items was omitted in listening in grade cluster 3–5 (0.3%).

AIR also conducted Rasch/IRT analysis of item responses from the 2004 and 2005 field test administration of the ELDA. AIR applied Master's (1982) partial-credit

RETURN to TABLE of CONTENTS

model to estimate ELDA item parameters for both multiple choice items responses and constructed response items. To evaluate item fit, both Infit and Outfit statistics were examined. Items were flagged if the Infit or the Outfit values were less than .7 or greater than 1.3. This model estimates the difficulties of dichotomously scored multiple-choice items as well as the difficulties of the steps involved in the solution of graded response constructed response items. As a result of the Infit and Outfit analyses of the 2005 field test administration, misfitting items were flagged in each grade cluster and domain. The number of misfitting items flagged ranged from only 1 in the reading domain for grades 3–5, to 36 items flagged for misfit in the reading domain for the same grade cluster.

Mantel-Haenszel DIF analyses were conducted for multiple choice (dichotomous) and grade response (polytomous) items, to detect bias among individual test items for different subgroups of students. (See *Bias Review* for further information.)

Test reliability. Cronbach's alpha was determined for each domain by to test form, subgroup, and grade level. The yielded coefficients for the 2004 field test showed a high internal reliability with a range from .822 for Reading, Form B (grades 3–5, ELL-exited) to .992 in Speaking, Form B (grades 9–12, monolingual English-speakers). The 2005 field test showed Cronbach's alpha ranges from .76 for grades 3–5 in Reading, Form A (grades 3–5) to .95 for Reading, Forms A and C, and for Listening, Form B (all grades 9–12).

Test validity. C-SAVE conducted two content validity studies: one studied students' measured English language proficiency levels (derived from ELDA scaled scores from the 2004 field test) and compared these scores with teacher ratings of student proficiency in each assessment area, as well as with scores obtained from the Idea Proficiency Test. (Kopriva et al, 2004). In multi-trait/multi-method path model analyses, ELDA Speaking scores were found to be most closely associated with teacher ratings of student speaking proficiency, providing evidence of convergent validity of ELDA Speaking scores. Kopriva et al. concluded that the ELDA Speaking assessment functions as intended for students from different types of English as a second language (ESL) programs and different primary language groups, but found discrepancies about the functioning of the ELDA Speaking assessment for students from western European language groups. Additionally, latent class analyses indicated that the ELDA distinguished five levels of language proficiency as the design of the assessment intended. However, Kopriva et al cautioned that, at the highest levels of complexity and difficulty of the ELDA assessment, these measures may not provide precise enough information to be used exclusively for decisions about exiting English language learners from language development programs.

As part of the classical test item analysis conducted by AIR, constructed response items were evaluated on the basis that low polyserial correlations might indicate issues with construct validity. Constructed response items, for example, were flagged if the adjusted polyserial correlations were lower than .10. This is because near zero or negative-adjusted polyserial correlations may indicate a flawed scoring rubric, mis-ordering of response categories, reader difficulties interpreting the rubric, or that the item does not measure the construct of interest. For both multiple choice and constructed items, omission rates of greater than 15% were also flagged, since this suggests that these items caused confusion for test takers on how to respond to the item, confusion among raters about how to score the item, or that the item was too difficult. IRT analysis of constructed response items also revealed items with construct validity problems; the items were flagged if their DIF statistics fell into the "C" category for any group. A DIF classification of "C" means that the item shows significant DIF, and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. These items were flagged regardless of whether the DIF statistic favored the focal or referent group.

As described in the *Test Development Section*, AIR conducted expert bias reviews of items in the initial item development process. Additionally, AIR conducted differential item functioning (DIF) analysis on all items from the 2004 and 2005 field test to determine whether items showed bias across sub-groups. Three DIF analyses were performed for each item: 1) ELL Spanish-speaking students vs. all other ELL language groups; 2) ELL-exited students vs. monolingual English speakers; and 3) all ELL language groups vs. non-ELL students (monolingual English speakers and ELL-exited students). Using the Mantel-Haenszel and generalized Mantel-Haenszel procedures for DIF analysis, a dichotomous item was flagged if the DIF statistic was lower than 0.2 or higher than 0.9. A constructed response item was flagged and reviewed if its DIF statistic (calculated using the Mantel-

RETURN to TABLE of CONTENTS

Haenszel chi-square procedure) was less than 0.2 or greater than 0.15. Items were classified into three categories (A, B, or C) ranging from no DIF to mild DIF to severe DIF according to common DIF classification conventions. Overall, relatively few items were flagged for DIF across all ELDA test forms following the 2005 test administration, with the exception of reading test items in grades 6–8 and listening test items in grades 9–12. LEP-SCASS then reviewed the flagged items. While many of these flagged items were subsequently suspended, most were approved following the expert review.

Technical Report

Published by AIR (2005). Available on the Council of Chief State School Officers' website: http://www.ccsso.org/projects/ELDA/Research_Studies/

Idaho English Language Assessment (IELA)

Grade Cluster(s): K; 1–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Idaho (spring 2006)

Test Purpose

Designed to fulfill requirements outlined in Title III of The No Child Left Behind Act, the IELA is used in Idaho to assess English language proficiency and is used with other information to determine exit and reclassification. The IELA is not used for placement decisions. A separate *English Language Learner Placement Test* is used for this purpose.

Score Reporting

At each grade cluster, except K, there are two forms. One form is designed to assess English proficiency at the beginning level and the other form is designed to assess proficiency at the intermediate level and above. Within each grade cluster, results on the Level 1 (Beginning) form and Level 2 (Intermediate) form are reported on the same scale. Scaled scores are provided in the areas of listening, speaking, reading, writing, and comprehension. The comprehension score is a composite of selected reading and listening items. A total scaled score is based on performance in all four language domains. Test performance in each language domain and comprehension is reported at three levels of proficiency: Beginning, Advanced Beginning to Intermediate, and Early Fluent and Above. Total IELA performance is reported at five levels of English proficiency: Beginning, Advanced Beginning, Intermediate, Early Fluent, and Fluent. A formal standard setting, conducted in August, 2006, established the correspondence between test performance and English proficiency level.

Test Development

IELA is a revised version of the Mountain West Assessment (MWA) developed by the Mountain West Assessment Consortium (MWAC). Please refer to the summary entry on the MWAC of this chapter for more information on initial test development. After the consortium disbanded in 2005, Idaho continued the test development with Questar Assessment, Inc. (formerly Touchstone Applied Science Associates [TASA]). In 2006, the level 1 (Beginning) and level 2 (Intermediate) forms at each grade cluster were linked by inserting a set of common items. A second set of forms, also based on the initial MWAC Assessment, was developed and administered in 2007 and subsequently equated to the 2006 forms.

A formal standard setting was conducted in 2006. Questar facilitated two panels of Idaho educators: one panel focused on test forms for Grades K-5, while the second panel considered test forms for middle and high school students. Cut scores were determined using the Bookmark or item mapping procedure. Panelists were given anonymous feedback about group recommendations after each round of deliberations. In addition, they were presented with statewide impact data following the second round.

Alignment to State Standards

The Idaho English Language Development (ELD) Standards were revised in 2006 under contract to WestEd. In September, 2006, a study was conducted by Assessment and Evaluation Concepts, a subsidiary of Questar Assessment, to determine the alignment of the IELA to the revised ELD Standards. This alignment study prompted a round of new item development to address those areas in which IELA was not well aligned with Idaho ELD Standards. New items that survive content and bias reviews and field testing will be incorporated into subsequent operational forms.

Technical Properties of the Test

Item analysis. Please see the summary in the entry on the Mountain West Assessment in this chapter for more information on initial item analysis conducted for this assessment. Data from the 2006 and 2007 administration of the IELA were analyzed using classical test (CT) and item response theory (IRT) methods.

Test reliability. Reliability is reported in terms of coefficient alpha and the Standard Error Measurement (SEM) which are based on the 2006 IELA test administration and calculated for each language domain and the Total IELA by grade. Alpha coefficients for total scores within each test form were consistently high, ranging from .85 to .96. Additional reliability information will be available after the spring 2007 test administration.

Test validity. Evidence for content validity of the IELA consists of initial benchmarking studies conducted by the MWAC and the subsequent study of test alignment to Idaho ELD Standards.

Evidence for criterion-related validity is included in the IELA Technical Report. Additional validation research is currently underway.

Please refer to the summary in the entry for the Mountain West Assessment in this chapter for more information on initial freedom from bias analyses.

Technical Report

Idaho State Board of Education (2006). *Idaho English language assessment IELA*. (Tech. Rep. 2006). Retrieved March 13, 2007, from http://www.boardofed.idaho.gov/lep/documents/06IELA-TechnicalReport-FINAL.pdf

IPT® Title III Testing System (IPT)

Grade Clusters: Pre-K; K; 1-2; 3-5; 6-8; 9-12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Date): Alaska (2006), North Carolina (2005)

Test Purpose

The new IPT® Title III Testing System (IPT) may be used by states to determine placement, progress and redesignation of ELLs, and for Title III reporting purposes.

Score Reporting

Standards scores are given in reading, writing, listening and speaking. In addition, a standard score in comprehension, which is a composite of listening and reading scores is given. Student performance across all four domains is used to give an overall English Proficiency standard score. Cut scores for reporting test results on state-specific English language proficiency levels were determined by individual states that adopted the IPT for their statewide English language proficiency assessment to meet the Title III requirements under NCLB.

Test Development

The new IPT® Title III Testing System was developed by Ballard and Tighe specifically for Title III compliance. Field testing for Form A occurred during the spring of 2004 and field testing for Form B took place in the spring of 2005. A pilot test of Form A was conducted in fall 2004 and a pilot test of Form B took place in fall of 2005. The assessment became operational in 2005 when North Carolina began using the test for identification and placement. Both Alaska and North Carolina used the new IPT for annual assessment in the spring of 2006.

Alignment to State Standards

An alignment study between the Alaska ELP standards and the new IPT® Title III Testing System was conducted in the spring of 2006 using Gary Cook's ELP application of Norman Webb's Web Alignment Tool. The alignment was commissioned by the Alaska department of Education and Early Development (EED), and submitted to EED in the spring of 2006. The state of North Carolina is in the process of commissioning an independent alignment study.

Standard Setting

Ballard & Tighe conducts standard setting using the item mapping method. Two standard setting studies have been conducted for North Carolina. The first workshop was conducted during the spring of 2005. These initial cut scores for the IPT were adopted by the North Carolina State Board of Education in November of 2005. A follow-up study was conducted in the spring and summer of 2006 to finalize the cut scores taking more recent impact data and experience with IPT test administration into account. The standard setting workshop for Alaska was conducted during the summer of 2006. The new IPT® Title III Testing System is vertically and horizontally scaled to allow for comparisons across test forms and levels.

RETURN to TABLE of CONTENTS

Technical Properties

Item Analysis. Item analysis was conducted using Rasch measurement. Multiple-choice items and 0-1 scored constructed response items were analyzed using the dichotomous Rasch model, while rubric-scored items were analyzed using the polytomous Rasch model. As part of the analysis, model fit and DIF statistics were evaluated for each item. For more information, see the Technical Manuals.

Test Reliability. Reliability for the IPT has been assessed through Classical Test Theory statistics and through score-specific IRT standard errors of measurement. Coefficient alpha was calculated by form (A or B) and grade span (K; 1-2; 3-5; 6-8; 9-12). Across grades in Form A, the coefficient alpha for the listening section ranges from 0.83 to 0.87. For speaking, the range is from 0.83 to 0.93. For reading, it is from 0.85 to 0.91. For comprehension, the alphas range between 0.89 and 0.94. Finally, the alphas range from 0.93 to 0.96 for the overall test. Across grades in Form B, the alpha ranged from 0.82 to 0.92 in listening, 0.89 to 0.93 in speaking, 0.83 to 0.94 in reading, 0.83 to 0.92 in writing, 0.88 to 0.95 in comprehension, and 0.94 to 0.97 for the overall score.

To determine inter-rater reliability, intra-class correlation coefficients were computed for written constructed-response items that are graded using rubrics. For Form A, these coefficients ranged from 0.75 to 0.93. For Form B, they ranged from 0.84 to 0.96. Please see the technical manual for further information on reliability analyses for this assessment.

Test Validity. The IPT technical manuals contain evidence of content validity, criterion-referenced validity, and construct validity for the tests [in accordance with the AERA, APA &NCME 1999 standards, Content validity was established across grade spans and forms by ensuring that the test content and administration procedures are developmentally appropriate. In grades 1-12, an additional content validity criterion was to ensure that test content focuses on academic English as defined by Cummins (2000), Bailey and Butler (2002) and Chamot and O'Malley (1994). In addition, to determine whether test scores were affected by construct irrelevant factors, the test developers qualitatively documented field and pilot test administrations to analyze the test interaction between students and the test administrator as well as the content of students' answers to constructed response items.

Construct validity was established through comparing scores on the test to other measures which intend to measure the same construct. Teacher opinions about students' abilities across domains and ability levels as well as standard scores from the IPT were compared using Analysis of Variance (ANOVA) for Forms A and B. Additionally for students in California taking the Form A test, CELDT scores were also used in the analysis. Criterion related validity could not be explored through direct means at the time of the publication of the technical manual. Therefore it was preliminarily explored through computation of ANOVA statistics comparing overall English standards scores and teacher ratings of students' ability.

Test development included a bias review of individual test items. In this process, the bias reviewers responded to a set of questions regarding each item and provided feedback about acceptability and recommendations for change. Additionally, statistical bias (DIF) statistics were computed for each item developed for the IPT using up to 20 different reference groups, depending on the data obtained during field and pilot testing of the IPT. Items were checked for DIF with respect to students' primary language, country of origin, gender, ethnicity, disability, and economic status. Ballard & Tighe maintains an item bank listing all computed DIF statistics for all items ever tested during the development of the IPT. Please see the technical manual for more information on freedom from bias analyses.

Technical Manuals

Ballard and Tighe (2006). IPT® K Technical Manual: Grade K, Form A. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® K Technical Manual: Grade K, Form B. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 1-2 Technical Manual: Grades 1-2, Form A. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 1-2 Technical Manual: Grades 1-2, Form B. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 3-5 Technical Manual: Grades 3-5, Form A. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 3-5 Technical Manual: Grades 3-5, Form B. Brea, CA: Ballard and Tighe, Publishers.

RETURN to TABLE of CONTENTS

Ballard and Tighe (2006). IPT® 6-8 Technical Manual: Grades 6-8, Form A. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 6-8 Technical Manual: Grades 6-8, Form B. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 9-12 Technical Manual: Grades 9-12, Form A. Brea, CA: Ballard and Tighe, Publishers.

Ballard and Tighe (2006). IPT® 9-12 Technical Manual: Grades 9-12, Form B. Brea, CA: Ballard and Tighe, Publishers.

IPT® 2004: IPT Early Literacy Test reading and writing (IPT Early Literacy Test, IPT Early Literacy R & W)

Grade Cluster(s): K–1[1]
Domains Tested: Reading and writing
Date(s) Published: 2004
State(s) Using This Test for Title III Accountability (Implementation Date):
Massachusetts (spring 2007)

Test Purpose

The IPT Early Literacy Test was designed to assess the literacy development of students in the domains of reading and writing. While one or two states have reported standard scores earned on the IPT Early Literacy Test as part of its language proficiency assessment for English learners in Grades Kindergarten through 2, this instrument was not developed as an achievement test, and it should not be used as the only measure of a student's English reading and writing proficiency. The IPT Early Literacy Test scores have been used to report standard scores of students.

Score Reporting

Standard scores, percentile ranks, and normal curve equivalent scores and ordinal reading and writing stage designations are provided. For information about IPT® 2004 in general please refer to the IPT® 2004 technical manual. It should be noted, however that the IPT® 2004 test exists apart from its use in Massachusetts. In particular, in the Score Reporting section, Massachusetts did report standard scores, but percentile ranks and normal curve equivalent scores were not provided.

Test Development

After test items were written, they underwent expert review. A field test of the IPT 2004 Early Literacy Test was conducted in spring 2000. The operational test was first published as the IPT Early Literacy Test in 2001. In 2004, the test name was changed to the IPT 2004 Early Literacy Test. In connection with that change, test norms were updated and a second edition of the test was created.

During the field testing, teachers were asked to classify students' oral English language proficiency into one of the following four categories: Non-English Speaking, Limited English Speaking, Fluent English Speaking and English-only (i.e., native English Speaking). These classifications were used to determine cut scores in grades K–1 for each of three reading stages: Pre-Reader, Beginning Reader, and Early Reader. These classifications were also used to determine cut scores in grades K–1 for each of the three writing stages: Pre-Writer, Beginning Writer, and Early Writer. In addition, Cramer's *V* and Pearson's *R* were calculated to assist in the determination of cut scores for each stage in both reading and writing.

Alignment to State Standards

The alignment of the test to content standards has not been formally studied.

Technical Properties of the Test

Item analysis. Item analysis was conducted using classical test theory statistics, yielding *p*-values. The mean *p*-value for the total reading test was .75 in kindergarten and .86 in first grade. For more information on item analysis, refer to the IPT® 2004 technical manual.

Test reliability. Two classical test theory measures were used to determine test reliability: Cronbach's alpha coefficients and the standard errors of measurement (*SEM*). The alpha of the total reading test was .89 in kindergarten and .90 in first grade. In addition, a study was conducted in a sample of kindergarten and first-grade students to determine inter-rater reliability. Pearson's *R* was provided as a measure of agreement between raters. In reading, Pearson's *R* was .868 in kindergarten and .850 in first grade, indicating strong evidence of inter-rater reliability. In reading, Pearson's *R* was .478 in kindergarten and .756

[1] The IPT-1 test was also given to all LEP students in grade 2 in spring 2007

RETURN to TABLE of CONTENTS

in first grade, indicating moderate-to-strong agreement among raters. All Pearson's *Rs* were significant at the .01 level. Please refer to the technical manual for more detailed information on test reliability.

Test validity. Intercorrelations between IPT Early Literacy reading domains were provided as evidence of the construct validity of the reading assessment. Correlations between the IPT Early Literacy Test in reading and teachers' opinions of student academic ability were included as measures of criterion-related validity. Overall moderate correlations of .1146 to .5100 were found across both grade levels. A content validity study for the reading test could not be located.

Bias review of items was conducted by content experts during the initial review of the test items. Please refer to the technical manual for additional detailed information on bias review.

Technical Reports

Ballard & Tighe (2006). IPT® 2004 *Technical manual: IPT® early literacy reading & writing grades K–1*. Brea, CA: Ballard & Tighe.

Kansas English Language Proficiency Assessment (KELPA)

Grade Cluster(s): K–1; 2–3; 4–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Kansas (spring 2006)

Test Purpose

The purpose of the KELPA is to assess annual progress of English learners in Kansas public schools and for reclassification to fluent English proficient.

Score Reporting

A scale score is given in each domain tested; an overall composite score is provided based on a differential weighting system by grade level. Kindergarten and first grade students have a higher weight placed on the domains of listening and speaking (30% to 35%) while second to twelfth grade students have more weight contributed by reading and writing.

In addition, for students taking the 2-3, 4-5, 6-8, and 9-12 grade level assessments, the writing domain score was assessed using two types of item formats. Students were asked to respond to 1) open-ended constructed response items to which they were asked to write to a choice of prompts and 2) a set of multiple choice items. Each of the two types of item formats was assigned a weight that was then used to calculate the domain score for writing. For all second through twelfth grades students, the open-ended writing performance comprised 50% of the writing domain score while the multiple choice contributed the remaining 50%.

Cut-scores for each of the four domains and the composite total score were determined by the KSDE based on information gathered using school-based content experts' item judgments, teacher ratings of student classroom performance, student performance on the state's general reading assessment tests, and the recommendations of teachers, curriculum directors, and principals reviewing the data. Based on the four weighted domain scores students are placed in one of four proficiency levels: Beginning, Intermediate, Advanced, and Fluent.

Test Development

The KELPA was developed specifically for Title III compliance as a result of collaboration between The Center for Educational Testing and Evaluation (CETE), the University of Kansas and the Kansas State Department of Education (KSDE).

Committees including ESL teachers and directors, content specialists and higher education worked with the Kansas State Department of Education ESOL consultant worked and the CETE staff to develop assessment items. The KELPA was field tested across the state in spring of 2005; revisions were implemented and field testing occurred again in fall of 2005 with full implementation in spring of 2006.

A committee of field practitioners and content experts convened to review KELPA items using a modified Angoff method and to determine recommended cut scores. Two separate committees convened to make recommendations for cut scores based upon KELPA data, teacher judgment and the state content reading assessment. The subsequent cut scores were adopted by the KSDE in August of 2006. Parallel forms of the KELPA within a grade band were made comparable using a common scale equating procedure.

RETURN to TABLE of CONTENTS

Alignment to State Standards

The English to Speakers of Other Languages (ESOL) Standards, adopted by the KSDE, served as the basis for the development of this assessment. The KSDE reports that an alignment study was conducted with the state-adopted ESOL standards as part of KELPA test development. A small committee including field representation compared each item on the 2005 version of the KELPA with the Kansas ESOL Standards to determine gaps. This resulted in changes reflected in the 2006 assessment.

Technical Properties of the Test

Item analysis. Item analysis procedures and findings could not be determined from the available information on the KELPA.

Test reliability. Information on test reliability studies could not be determined from the available information on the KELPA.

Test validity. Information on test validity and freedom from bias could not be determined from the available information on the KELPA.

Technical Report and Administration Manuals

A technical report could not be located at the time of publication.

Language Assessment Systems Links (LAS Links)

Grade Cluster(s): K–1; 2–3; 4–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: LAS Links Form A (2005) and LAS Links Form B (2006)[2]
States Using This Test for Title III Accountability (Implementation Date):
Connecticut (winter & spring 2006), Hawaii (spring 2006), Indiana (winter & spring 2006), Maryland (spring 2006), Nevada (2005–2006 academic year)

Test Purpose

Developed to satisfy the requirements of Title III of the No Child Left Behind Act, LAS Links is also used to:

- provide placement information, ongoing information on student growth and summative information on students' acquisition of English
- meet federal and state testing requirements
- provide information that educators may use to improve instruction.

Score Reporting

Scores are provided in reading, writing, listening, and speaking. An oral domain score is derived from the listening and speaking domain scores. A comprehension score, derived from listening and reading domains, is also provided. Based on assessment results the titles of the five performance levels may be different among the states: Beginning, Early Intermediate, Intermediate, Proficient, and Above Proficient. However, the state of Maryland uses the following performance levels: high beginning, low beginning, low intermediate, high intermediate and advanced. The cut scores for these levels in Maryland were derived after a cut score review process.

Test Development

LAS Links was developed by CTB/McGraw Hill in response to Title III requirements. The test blueprint for LAS Links was based upon the English Language Proficiency Framework developed from language acquisition models, the National Teachers of English to Speaker of Other Languages (TESOL) Standards and upon several states' standards for English language development (ELD). Social and academic English were considered in the development of objectives and test items within the domains. For more information on item development and item review please refer to the technical manual.

Fifteen LAS Links forms (three test forms for each of five grade cluster levels) were field tested with students in California, Florida, New York, Texas and Washington as well as Brazil, Chile, China, India, Jordan and Mexico[3]. LAS Links was first used across grades in Colorado, Connecticut, Hawaii, Indiana, Maryland, and Nevada in 2006. The State of Colorado contracted with CTB/McGraw Hill to develop the Colorado English Language Assessment (CELA). The CELA is closely aligned to LAS Links.

Standard setting was led by CTB/McGraw Hill in June 2005. A modified bookmark approach was used by national panel of educators of English language learners. Common scaling was utilized in the development of this assessment. Additional information on scoring and standard setting is located in the technical manual.

[2] The two forms are parallel forms and include all the grade spans and domains of reading, writing, listening, and speaking.

RETURN to TABLE of CONTENTS

Alignment to State Standards

Alignment studies were conducted comparing LAS Links to the adopted content standards of several states, including Nevada and Connecticut. In addition, CTB/McGraw Hill conducted two types of alignment studies between the National TESOL standards and states' ELD/ELP standards. One alignment study conducted a document analysis by content experts. A second alignment study applied the modified Webb's (1997) alignment model to ensure depth and breadth of alignment with the TESOL standards and with various states' ELD/ELP standards. The LAS Links Technical Manual (2006) reports the findings of the document analysis and the correlation between LAS Links domain objectives with TESOL goals and standards. Specific information regarding the results of alignment studies between specific states' standards could not be located in the technical manual.

Technical Properties of the Test

Item analysis. Classical test theory (CTT) methods were also used to determine item difficulty through the calculation of *p*-values. Items were also analyzed using an item response theory (IRT) application. Item difficulty statistics are provided for speaking, reading, writing, comprehension and oral domain items by grade span and by item. Average difficulty by grade span is also provided. Raw score descriptive statistics for field test Forms A and B are also provided for both multiple choice and constructed response items. Refer to the technical manual for specific information on item analysis.

Test reliability. In order to ensure reliability of constructed response items, check sets, read behinds, and double-blind reads were implemented. Cronbach's alpha-alpha coefficients ranged from .78 in both Listening Form B (K–1) and Reading Form B (kindergarten) to .95 in Speaking Form A and B (grades K–1), and Oral Form A (K–1) and Oral Form B (grades 2–3). In addition, the following methods were used as measures of reliability:

- classical standard error of measurement (*SEM*)
- conditional standard error of measurement based on item response theory
- intraclass correlation coefficients (to evaluate inter-rater agreement)
- weighted kappa coefficients (to measure reader agreement)
- IRT methods (to create test characteristics curves)

Please refer to the technical manual for detailed information on these reliability measures.

Test validity. Test content validity was addressed in the test development process by educational experts to determine level of alignment with instructional goals. Items were reviewed to ensure that items align to subject matter.

DIF analysis determined test validity. Items which displayed poor item statistics or differential item functioning (DIF) were excluded or given lower priority in item selection when items were developed.

DIF analysis of test items also compared test item parameters across gender. Because of the relatively small sample size, the Linn and Harnisch procedure (1981) was used with dichotomous test items included in the pilot version of LAS Links. A generalization of the Linn and Harnisch procedure was also used to measure DIF for constructed-response items. Please refer to the technical manual for further information on the DIF analysis.

Item developers for LAS Links used the following guidelines to minimize test bias: *Guidelines for Bias-Free Publishing* (MacMillan/McGraw-Hill, 1993a) and *Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (Macmillan/McGraw-Hill, 1993b). Additionally, there were internal bias reviews of LAS Links assessment materials. Thirdly, educational community professionals representing various ethnic groups reviewed pilot materials for possible bias in language, subject matter, and representations of diversity. Please refer to the technical manual for further information on validity studies.

Technical Reports

CTB/McGraw-Hill (2006). *LAS links technical manual.* Monterey, CA: CTB/McGraw-Hill LLC.

Maculaitis Assessment of Competencies II (MAC II)

Grade Cluster(s): K–1; 2–3; 4–5; 6–8; 9–12
Domains Tested: Reading, writing, listening and speaking
Date(s) Published: 2001
State(s) Using This Test for Title III Accountability (Implementation Date): Missouri (February 2002). Missouri is now in the process of revising

[3] Field testing dates could not be located in the technical manual.

RETURN to TABLE of CONTENTS

their standards. After their contract for the MAC II expires in 2008, they may be working with the publishers of MAC II to customize the test according to the state's needs.

Test Purpose

The MAC II was created to identify students, inform student placement monitor student progress, inform instruction, determine program exit, and aid in program evaluation.

Score Reporting

Scaled scores are provided for each domain, listening, speaking, reading, writing, and comprehension as well as for the total test. For each domain, students are assigned one of five competency levels: Basic Beginner, Beginner, Low and High Intermediate, and Advanced. The reading domain for grades 4 and up also gives a criterion referenced score which is reported on the Degrees of Reading Power (DRP) readability scale. The test includes grades K-12 national norms for each language domain, the total test, and DRP Reading comprehension where applicable.

Test Development

The MAC II test is a revised and updated version of the MAC originally published in 1982. The test was acquired by Touchstone Applied Science Associates, Inc. (TASA)[4] in 1997 and revised. Items were field tested from 1999-2000. A national research program was undertaken by TASA during 2000 and 2001 to establish norming data and to determine the reliability and validity of the test.

A modified Angoff procedure was used to determine cut scores for proficiency levels within each domain. External criteria such as the students' level in ESL instructional programs, the estimate of English language proficiency, and a general assessment of students' academic performance were used to determine a range of standard scores by English level competency within each domain. Cut scores indicating overall level of performance on the MAC II were established for English proficiency using a similar procedure. However, the comparison of distribution of scores for ELLs and native English speakers was weighted more heavily in the cut scoring process. Specific information on scoring and standard setting is available in the technical report.

Alignment to State Standards

The MAC II content is based on National TESOL standards, but is not aligned to particular state ELL or content standards.

Technical Properties of the Test

Item analysis. Test items were analyzed with procedures from Classical Test Theory (CTT) and Item Response Theory (IRT) approaches. CTT analyses included P-values, point-biserial correlations, and the distribution of responses among distracters. The Rasch (IRT) model was used for dichotomous items and the Partial Credit Model for polytomous items. Please refer to the technical report for further information.

Test reliability. Internal consistency reliability coefficients and raw score standard errors of measurement were calculated for each domain and for the total test. Reliability coefficients ranged from .97 to .79. For constructed response items, inter-rater reliability showed that scorers were in agreement on 73.1 percent of the items and were within one point on 95.7 percent of the items. Specific information on reliability is available in the technical manual.

Test validity. To ensure content validity, a panel of English as a Second Language (ESL) and bilingual teachers reviewed the test for clarity, potentially confusing items, and age appropriateness. They also reviewed the directions to students and administrators in order to suggest changes to these and any other design aspects of the test.

The construct and criterion-related validity of the test was determined through the following methods:

- Correlation validity–Pearson product-moment correlations of the individual tests of the MAC II and performance on the DRP portion of the reading test was conducted by test level in grades 4–12. Pearson product-moment correlation was also used to compare performance on the MAC II to other published tests of English proficiency, including the Language Assessment Scales (LAS) writing test, the Stanford Diagnostic Reading test (SDRT4) reading comprehension and vocabulary tests, the Secondary Level English Proficiency Test (SLEP) reading and listening test, and the IDEA IPT II Oral Language Proficiency Tests. The strongest correlations between the MAC II and these other assessments were .77 and .76 (MAC II Speaking and Listening with the IDEA-IPT II);

[4] Now Questar Assessment, Inc.

RETURN to TABLE of CONTENTS

.67 (MAC II Reading with the SLEP Reading) and .77 (MAC II Writing with the LAS Writing domain).

- The relationship of students' performance on the MAC II to teacher rating of English language proficiency and overall academic proficiency and the student's program placement was examined. Specific information on validity is provided in the technical report.

Freedom from bias. The panel of ESL and bilingual teachers also reviewed items for freedom from bias. Material which might be considered offensive to particular cultural groups or which suggested stereotypes were revised or dropped. Specific information on freedom from bias is available in the technical manual.

Technical Reports

Maculaitis, J.D. (2003). *The MAC II test of English language proficiency handbook with norms tables A and B test forms.* Brewster, NY: Touchstone Applied Science Associates

Massachusetts English Language Assessment-Oral (MELA-O)

Grade Cluster(s): K–12
Domains Tested: Listening and speaking
Date(s) Published: 2003
State(s) Using This Test for Title III Accountability (Implementation Date):
Massachusetts (2004)[5]

Test Purpose

The MELA-O is used for review of English learners' progress in listening and speaking skills for state and federal accountability purposes.

Score Reporting

The MELA-O uses an observation protocol in which a scoring guide is used to rate the student's levels of Comprehension (listening) and Production (speaking). Individual students are observed in regular classroom activities over a month-long assessment window. A score from 0 to 5 is determined for Comprehension (listening) and for each of the following subdomains of Production (speaking): Fluency, Vocabulary, Pronunciation, and Grammar. All reported scores are raw scores. Based on the results of both the MELA-O and MEPA-R/W, individual students are assigned one of four overall proficiency levels—Beginning, Early Intermediate, Intermediate, and Transitioning.

Test Development

The MELA-O was developed by the Massachusetts Department of Education, in collaboration with the Massachusetts Advisory Group and the Evaluation Assistance Center (EAC) East. The EAC East was led by the Center for Equity and Excellence at George Washington University between 1991 and 1995. The MELA-O scoring matrix is based on the Student Oral Language Observation Matrix (SOLOM) and the Student Oral Proficiency Rating (SOPR). It was developed, piloted, and field-tested between 1992 and 1995. The first operational administration of the MELA-O for Title III reporting purposes occurred during the 2004-2005 school year.

For more information on standard-setting process please refer to the MEPA Technical Report.

Alignment to State Standards

The MELA-O scoring matrix was developed to align with the listening and speaking skills outlined in *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (2003) adopted by the Massachusetts Department of Education.

Technical Properties of the Test

Item analysis. MELA-O is an observation protocol.

Test reliability. Test developers determined test reliability by conducting the following studies:

- A pilot study examining inter-rater reliability was undertaken by the EAC East from 1993 and 1994. The study found that inter-rater reliability for the MELA-O was .74 in listening domain and .75 in the speaking domain. A second pilot study undertaken in 1994–1995 found that the inter-rater reliability was .77 in the listening domain and .81 in the speaking domain.
- Kappa coefficients were also used to determine inter-rater reliability.

[5] The MELA-O was administered during fall and spring 2004–2005 to all English language learners (ELLs) in grades K-12. In fall 2006, it was administered to all ELL students in kindergarten, all ELL students in grade 3, and those ELL students in grades 1–2 and 4–12 who did not participate in the spring 2006 MELA-O test administration. In spring 2007, all ELL students in K–12 as well as all former ELL students in grades K–12 will take the exam.

RETURN to TABLE of CONTENTS

- A split-half reliability study was conducted for the MEPA (the combined MELA-O and MEPA-R/W), yielding Cronbach's alpha coefficients to compare individual student performance across test administrations.
- Stratified coefficients (based on test item types) were calculated for each grade span, administration, and combination of sessions taken to determine internal reliability.

In addition, test reliability was analyzed using standard error of measure, descriptive statistics of the composite MEPA scores, and test characteristic curves (TCC) using IRT methods.

Test validity. Information on test validity and test bias analyses for the MELA-O could not be located prior to publication of this report.

Technical Report

Massachusetts Department of Education (2005). *2005 MEPA technical report.* Retrieved on July 1, 2007, from http://iservices.measuredprogress.org/MEPA%20Report%20main%20body%20%20Final.pdf.

Massachusetts English Proficiency Assessment–Reading & Writing (MEPA-R/W)

Grade Cluster(s): 3–4; 5–6; 7–8; 9–12
Domains Tested: Reading and writing
Date(s) Published: 2004
State(s) Using This Test for Title III Accountability (Implementation Date): Massachusetts (spring 2004)

Test Purpose

The primary purpose of the MEPA-R/W is to measure the progress of ELL students in acquiring proficiency in reading and writing in English. It is used in Massachusetts in combination with the MELA-O for state and federal accountability purposes. The state does not use currently use the MEPA-R/W to make program exit decisions.

Score Reporting

Prior to the test administration, each English learner[6] is assigned to two of three reading sessions. For reading and writing test administrations, Sessions 1 and 2 assess Beginning and Early Intermediate reading and writing performance, while Sessions 2 and 3 assess Intermediate and Transitioning reading and writing performances. Placement decisions for test sessions are determined at the school level and are based upon prior English language assessments, classroom observations and school work. Item difficulty associated with the session is factored into the final scoring of the reading and writing subsets.

Scaled scores from the reading and writing domains of the MEPA-R/W and raw scores from the MELA-O for listening and speaking domains are used to determine the overall MEPA scaled score. A statistical formula is used to map the total of the MEPA-R/W scaled score with the total MELA-O raw score in order to calculate each student's overall MEPA scaled score. Based on students' average scaled score across all domains tested, students are placed in one of four performance levels: Beginning, Early Intermediate, Intermediate, and Transitioning.

Test Development

The Commonwealth of Massachusetts, local educators, and Measured Progress developed the MEPA-R/W. Information on item development and field testing was not available. The MEPA-R/W was first administered during the 2004–2005 academic school year. In fall 2005, the MEPA was also administered to ELL students in grades 3–12 who did not have a baseline score from the spring 2005 MEPA administrations. In March 2006, 31,842 LEP students in grades 3–12 took the MEPA-R/W. The Massachusetts Department of Education used the IPT Early Literacy tests and IPT-1 tests to assess LEP students in grades K-2 beginning in spring 2007, and will develop customized K-2 MEPA-R/W tests in the next 1-2 years. The Massachusetts Department of Education conducted standard-setting sessions using the Body of Work Method to find the minimum score required to attain each of the program's performance levels.

Alignment to State Standards

The MEPA-R/W was created based on Massachusetts' *English language Proficiency Benchmarks and Outcomes for English Language Learners*, which were adopted by the Commonwealth of Massachusetts in 2003. The

[6] Massachusetts refers to English learners as LEP students, or "limited English proficient" students.

Massachusetts English Language Arts Curriculum Framework serves as the primary foundation for the Massachusetts' *English Language Proficiency Benchmarks and Outcomes for English Language Learners*. An alignment study could not be located.

Technical Properties of the Test

Item analysis. Classical test theory statistics and IRT models were used to determine item functioning. The logistic form of the one-parameter partial credit model was used for polytomous items.

Test Reliability. Please see the summary for the MELA-O for information regarding reliability studies for the MEPA-R/W.

Test validity. Information on validity or bias review could be located prior to publication of this report.

Technical Report

Massachusetts Department of Education (2005). *2005 MEPA technical report.* Retrieved on July 1, 2007, from, http://iservices.measuredprogress.org/MEPA%20Report%20main%20body%20%20Final.pdf

Michigan English Language Proficiency Assessment (MI-ELPA)

Grade Cluster(s): K–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Date): Michigan (2006)

Test Purpose

The MI-ELPA is used:

- to monitor annual ELL progress in acquiring English language proficiency,
- to determine exit from the ESL or bilingual program,
- to provide targets of proficiency for students to meet in each of the four domains tested, and
- for program placement, using a shorter version of the test.

Score Reporting

Scores are provided in each domain, reading, writing, listening, and speaking, as well as a total test score. The comprehension score is a composite of the listening and reading scores, while the total test score is an aggregate of the reading, writing, listening, speaking, and comprehension scores. Based on test performance, students are placed into one of four performance levels: Beginning, Intermediate A, Intermediate B, and Proficient.

Test Development

The Michigan ELPA was developed by Harcourt Assessment, Inc. and the Michigan Department of Education. For the spring 2006 test administration, the Michigan ELPA used items from the Harcourt Stanford English Language Proficiency assessment (SELP), items developed by the Mountain West Assessment Consortium, and items from the Michigan Educational Assessment Program (MEAP).

In order to determine cut scores, standard setting for the Michigan ELPA was conducted in July 2006 by Assessment and Evaluation Services in collaboration with Harcourt Assessment, Inc., and using an expert panel. The item-mapping/bookmarking procedure was used for standard setting. The MI-ELPA is a vertically scaled assessment.

The 2006 administration also included new embedded field test items. With the 2007 test administration of MI-ELPA, newer field tested items and fewer Harcourt and Mountain West Assessment items will be used.

Alignment to State Standards

The ELPA is aligned to Michigan's English language proficiency standards. Assessment specialists at Harcourt and ELL specialists reviewed the items on the 2006 operational forms to ensure that these items match the state's ESL standards. Specific item-mapping procedures were used in the test development process. Please see the technical manuals for more information regarding alignment studies.

Technical Properties of the Test

Item Analysis. Items on this assessment were analyzed through the classical test theory (CTT) and item response theory (IRT) frameworks. CTT analyses included calculation of *p*-values and point-biserial correlations. *P*-values and point-biserial calculations are reported in the technical manuals by grade-level cluster and by form for each item. IRT models used include the Rasch model for dichotomous items and the partial-credit model for polytomous items. Rasch difficulty, standard error of Rasch difficulty, INFIT, and OUTFIT were provided. Specific information on item analysis can be found in the technical manuals.

Test reliability. Cronbach's alpha by grade, classical standard error of measurement (*SEM*), conditional *SEM*, inter-rater reliability, reliability of each domain, and the reliability of classification decision at the proficient cut are provided to determine test reliability. Cronbach's alpha for the total test across grade levels (K–12) ranged from a low of .89 in kindergarten to a high of .96 in grade 9. In addition, Cronbach's alpha was provided by domain at each grade level cluster. Alpha ranged from a low of .70 in listening at the K–2 cluster to a high of .96 in speaking at the 9–12 grade cluster. Inter-rater reliability was assessed by determining the rate of agreement between readers' scores and team leaders' check scores on approximately 20% of the test booklets. The agreement rate between the readers' scores and the team leaders' score was 86%. Information on decision accuracy and consistency are provided by grade level in the technical manuals. Specific information on reliability is outlined in the technical manuals.

Test validity. Content, construct and criterion-related validity were examined for the MI-ELPA. Content validity was established during test development; item development activity ensured that items mapped to performance-level descriptors were based on Michigan's English language proficiency standards. Construct validity was established through the calculation of intercorrelations among domains by grade. Point-biserials and fit statistics are also offered as additional evidence of construct validity. Tests to establish criterion-related validity were only conducted on SELP items. See technical information on the SELP to see the specific tests conducted. Specific information on test validity is available in the technical report.

Assessment experts at Harcourt and ELL specialists reviewed the items from the Harcourt ELL item bank to ensure that the items were free from bias. In addition, differential item functioning (DIF) was performed on Mountain West Assessment and SELP items prior to administration. Additional embedded field-test items were analyzed for DIF using the Mantel statistic by comparing white and Hispanic students and males and females. Items were categorized as follows: "no-DIF" (A), "mild-DIF" (B) or "extreme-DIF" (C). Items showing moderate and extreme DIF were examined for bias. The standardized mean differences (SMD) were used as an effect type index for DIF. Specific information on freedom from bias is available in the technical report.

Technical Reports

Michigan Department of Education (2006). 12-05-2006 DRAFT *Spring 2006 English language proficiency assessment technical manuals*. San Antonio, TX: Harcourt Assessment, Inc.

Michigan Department of Education (2007). *Michigan English language proficiency assessment (MI-ELPA): Technical manuals appendix: 2006 Administration. Kindergarten through grade 12*. Retrieved July 2, 2007, from http://michigan.gov/documents/mde/MI-ELPA_Appendices_final_199605_7.pdf

Minnesota Modified Student Oral Language Observation Matrix (MN SOLOM)

Grade Cluster(s): K–12
Domains Tested: Lis tening and speaking
Date(s) Published: 2003
State(s) Using This Test for Title III Accountability (Implementation Date): Minnesota (2002–2003 academic year)

Test Purpose

The MN SOLOM is used to determine progress in the oral language domain only. It is sometimes used as one piece of information to determine exit from alternative instructional programs.

Score Reporting

The areas scored are listening (specifically academic comprehension and social comprehension) and speaking (specifically in the areas of fluency, vocabulary, pronunciation, and grammar). A five-point rating scale is utilized by the teacher for each student. The scores can be considered for individual domains or can be combined to give a total score. The range of the total score is between five and 30. A score of 22 or higher is considered to be Proficient and represents whether a student can participate in grade-level oral language tasks.

RETURN to TABLE of CONTENTS

Test Development

The MN SOLOM is an adaptation of the Student Oral Language Observation Matrix (SOLOM) by the San Jose Area Bilingual Consortium. Subsequently, this test has undergone revisions under the Bilingual Education Office of the California Department of Education. It is unclear when the SOLOM was first developed. The test is not copyrighted and can be copied or changed to meet local assessment needs.

Alignment to State Standards

Information on alignment of this test to state standards was not available.

Standard Setting

Information on scoring and standard setting was not available.

Technical Properties of the Test

Item analysis. Item analysis was not conducted.

Test reliability. Teachers using the MN SOLOM for review and exit purposes must undergo rater-reliability training. Additional information on test reliability was not available.

Test validity. Information on the validity of this assessment was not available.

Technical Reports

A technical report could not be acquired for this assessment.

Montana Comprehensive Assessment System English Proficiency Assessment (MontCAS ELP)

Grade Clusters: K; 1-2; 3-5; 6-8; 9-12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Montana (winter 2006)

Test Purpose

The MontCAS ELP is used for review of ELL progress and to provide Montana educators with proficiency scores for use in their schools, systems, and state.

Score Reporting

A score is given in reading, writing, listening and speaking, and comprehension. The comprehension score is a composite of reading and listening scores. A composite score is given to determine overall proficiency. Formal proficiency-level cut scores were not determined for the MWAC English proficiency assessments. Measured Progress led a group of national experts in English language acquisition in recommending cut scores for state panels.

Test Development

The MontCAS ELP was created in collaboration with Measured Progress and the Mountain West Assessment Consortium. The consortium initially consisted of the following states: Alaska, Colorado, Idaho, Michigan, Montana, Nevada, New Mexico, North Dakota, Oregon, Utah, and Wyoming. Items were developed by local specialists and educators from each member state. The test was first piloted in spring of 2004. Based on results from this initial field test, a new test was created and field tested in the fall of 2004. Following field testing, three final forms were developed in the winter of 2005. The MontCAS was first administered in Montana during winter 2006. Touchstone Applied Science Associates (now Questar Assessment) became the new test contractor for the MontCAS in the winter of 2007.

Alignment to State Standards

The assessment was initially designed to be aligned to Colorado's English language development (ELD) standards as a starting point to develop Mountain West Assessment Consortium ELD standards. According to the state of Montana, the MontCAS is aligned to state standards. An alignment study examining the test alignment to Montana's state standards could not be acquired.

Standard Setting

A Modified-Bookmark method for standard setting was used to recommend cut scores. Ultimately, however, it was recommended that states conduct standard setting using their own data. Additional information on scoring and standard setting could not be located.

Technical Properties of the Test

Item Analysis. Average difficulty and discrimination statistics are given for this assessment. In general the item difficulty and discrimination indices are within acceptable ranges. Please see the Mountain West Consortium chapter in this report for more information on item analysis.

Test Reliability. Information on the reliability of this assessment was not available.

Test Validity. Information on the validity of this assessment was not available.

Freedom from Bias. Bias and sensitivity reviews of all items to be piloted in spring 2004

were completed by a committee. Please see the Mountain West Consortium chapter in this report for more information on freedom from bias. Differential Item Functioning (DIF) could not be performed for this assessment due to limitations related to small sample size.

Technical Report

A technical report could not be acquired for this assessment.

Mountain West Assessment (MWA)

Grade Cluster(s): K; 1-2; 3-6; 7-8; 9-12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: N/A
State(s) Using This Test for Title III Accountability (Implementation Date): N/A

Test Purpose

The Mountain West Assessment was designed to meet the assessment guidelines within Title III of the No Child Left Behind act. Additionally, the Mountain West Consortium endeavored to design an instrument that could accurately assess the English proficiency progress of English learners in realistic academic contexts. With this in mind, consortium members endeavored to design an instrument as a tool for learning and for informing instruction, not simply for identifying and tracking English learners.

Score Reporting

Standard setting was not conducted for this assessment. However, two cut scores were recommended by the expert panel for grade span 3-5 and two for grade span 9-12. For each of these grade spans, the cut scores determined Emergent/Intermediate and Fluent/Advanced cut-scores. By applying an equipercentile smoothing technique, an average of the percentage of students above and below each cut-score was taken and applied to all grade spans. These recommended cut scores could be used by states who wished to develop an English language assessment instrument based upon the work of the MWAC.

Test Development

In 2003, the Mountain West Assessment Consortium (MWAC) received a two year Enhanced Assessment Grant from the U. S. Department of Education, with the Utah State Office of Education serving as the official agent to create the MWA. The consortium initially consisted of the following states: Alaska, Colorado, Idaho, Michigan, Montana, Nevada, New Mexico, North Dakota, Oregon, Utah, and Wyoming. Items were developed by local specialists and educators from each member state. The test was piloted in spring of 2004 for the first time. Based on results from this initial field test, a new test was created and field tested in fall of 2004. Following field testing, three final forms were developed in winter of 2005. The test instrument developed by the MWAC was not fully operational by the time the grant ended in 2005. However, three states have continued to develop this assessment and use some form of it: Idaho, Montana, and Utah. Idaho uses the Idaho English Language Assessment (IELA), Montana uses the Montana Comprehensive Assessment System English Proficiency Assessment (MontCAS) and Utah uses the Utah Academic Language Proficiency Assessment (UALPA). Please see summary entries for the above mentioned assessments in this chapter. Two other states, Michigan and New Mexico, use some items from the Mountain West Assessment in their English language proficiency assessments.

Standards Alignment

The assessment was initially designed to be aligned to Colorado's English language development (ELD) standards as a starting point to develop Mountain West Assessment Consortium ELD standards. These common standards were later referred to as the "Fountain Document." The test blueprint was then created using these ELD standards. Alignment studies for states using the Mountain West Assessment, however, were not conducted at the close of the grant.

Technical Properties of the Test

Item Analysis. Average difficulty and discrimination statistics are given for this assessment. In general the item difficulty and discrimination indices are within acceptable ranges. Please see the Mountain West Consortium chapter in this report for more information.

Reliability. Information on test reliability was not collected for this assessment.

Validity. Scope of work of the MWAC grant did not include validity studies.

Freedom from Bias. As part of the item development process, individual items and accompanying graphics were reviewed by state-selected participants for bias and sensitivity towards different ethnic, gender, cultural, or religious groups. Differential Item Functioning (DIF) could not be performed for this assessment due to limitations related to small sample size.

Technical Reports

A technical report could not be acquired for this assessment.

New Mexico English Language Proficiency Assessment (NMELPA)

Grade Cluster(s): K[7]; 1–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Compliance (Implementation Date): New Mexico (spring 2006)

Test Purpose

Besides being used for Title III accountability, the NMELPA is used to:

- measure annual progress of English language learners in acquiring and attaining English proficiency.
- focus educators on specific instructional standards that must be addressed in the classroom.
- NMELPA is not used for initial placement decisions. A short placement test, the New Mexico English Language Placement Test (NMELPT), is used for this purpose.

Score Reporting

One of five proficiency levels is administered for each of the four domains, based upon an individual's scale scores: Beginning, Early Intermediate, Intermediate, Early Advanced and Advanced. A composite score and overall proficiency level is determined, based upon the combined domain scores. The advanced level yields proficient status.

Test Development

The NMELPA is an augmented version of the Stanford English Language Proficiency assessment (SELP). New Mexico began development of its augmented version in early 2006 with Harcourt Assessment. The assessment uses a few items from the Mountain West Assessment Consortium item pool, the SELP, and the Stanford 9 assessment as well as items created by Harcourt Assessment specifically for the NMELPA. This assessment was not field tested; however, the test was implemented throughout the state in spring 2006. Item data analysis from this initial test administration was used by a bilingual review committee to make some necessary replacements for the 2007-2008 version. In addition, the reading and writing kindergarten domains were completely replaced with developmentally appropriate items from Harcourt's preliteracy SELP series. Beginning with school year 2007-2008, kindergarteners will be tested only during the spring testing window.

Harcourt Assessment led standard setting in Albuquerque, New Mexico, on June 19–21, 2006. New Mexico educators participated in the standard-setting process using the modified Angoff procedure. Please see the technical report for specific information on scoring and standards.

Standards Alignment

The New Mexico Public Education Department reports that the assessment is aligned with the state's ELP standards. Harcourt Assessment performed an evaluation of New Mexico's ELP standards and the SELP and found a consistent alignment, however some standards could not be easily assessed through a large-scale paper-and-pencil test, so these standards were not included on the test. In order to more closely align the assessment to New Mexico's standards, the test was augmented using items from the Stanford 9 assessment (Form T) along with a few items from the Mountain West Assessment Consortium item pool.

Technical Properties of Test

Item Analysis. Classical test theory (CTT) and item response theory (IRT) were used to analyze test items. The CTT included *p*-values, a measure of item difficulty, and point-biserial correlations, a measure of item discrimination. IRT models examined included the Rasch model for dichotomous items and the partial-credit model

[7] Beginning with the 2007-2008 school year, kindergarten will only be tested in the spring window.

RETURN to TABLE of CONTENTS

for polytomous items. Rasch difficulty, the standard error of Rasch difficulty, INFIT and OUTFIT were calculated. Please see the technical report for specific information on item analysis.

Test reliability. Reliability was estimated using Cronbach's alpha and the classical standard error of measurement (*SEM*). The alpha and *SEM* are provided by grade level/gender, grade level/special education, grade level/migrant status, grade level/immigrant status, grade level/placement in bilingual program, grade level/Title III status, grade level/length of enrollment, grade level/socioeconomic status, and grade level/ethnicity. Please see the technical manual for a table of values. The conditional standard error of measurement (*CSEM*), item response theory (IRT) statistics, reliability of classification decision at proficient cut, and inter- and intrarater reliability were also analyzed. See the technical manual for tables giving values of the *CSEM*, IRT statistics, the decision accuracy and consistency analyses for each of the four cut points that define the five performance levels on the NMELPA, and inter-rater reliability measures using the kappa statistic.

Test validity. As evidence of content validity, items were matched to align with instructional and state standards and reviewed to ensure adherence to standards. In order to investigate the internal structure of the assessment, correlations were obtained between the four domains. Intercorrelations ranged from .21 in kindergarten speaking and reading to .81 in grade 10 reading and writing. Across grades, correlations were highest between the reading and writing domains. The speaking domains were not as positively correlated to other domains, especially at the higher grades. A study was also undertaken to determine the relationship of the NMELPA to the state's regular assessment, the SBA. Data from this study provide evidence that the NMELPA is able to consistently distinguish ELLs whose English language proficiency is equivalent to that of non-ELLs from ELLs whose language skills are a barrier to learning in a traditional English-language classroom setting. Specific information on validity is located in the technical manual.

Freedom from bias. Harcourt assessment experts reviewed items on this test to ensure freedom from bias. Specific information on freedom from bias is available in the technical manual.

Technical Report

Harcourt Assessment (2007). *New Mexico English Language Proficiency Assessment Technical Manual*. New Mexico Department of Education. Retrieved August 16, 2007 from http://www.ped.state.nm.us/div/acc.assess/assess/dl/NMELPA/NMELPATechReportSpring2006.pdf

New York State English as a Second Language Achievement Test (NYSESLAT)

Grade Cluster(s): K–1; 2–4; 5–6; 7–8; 9–12.
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Date): New York (2005)

Test Purpose

Developed to comply with assessment requirements of Title III of the No Child Left Behind Act, NYSESLAT is also:

- designed to measure English learners' progress towards English language proficiency
- help schools to determine which standards needed to be addressed by teachers to assist English learners to be successful in the regular classroom
- gives bilingual and ESL teachers, valuable information to inform and adapt classroom instruction to meet identified needs of their students.

Score Reporting

Students are provided scores in listening/speaking and reading/writing. Based on test results, students are placed in one of four proficiency levels: Beginning, Intermediate, Advanced, and Proficient.

Test Development

Development of the NYSESLAT began as a joint effort between Educational Testing Service (ETS) and the New York State Education Department. Later, the state began to work with Harcourt Assessment Inc. to complete test development. For the 2005 test administration, items from the Harcourt English language learner item bank were initially used to construct the newly developed NYSESLAT.

Items from the Harcourt ELL item bank included items developed for the Stanford English Language Proficiency (SELP) test forms. Pre-writing items were developed and field tested in January 2005; these items were then used on the 2005 operational test form. New items were subsequently developed for the spring 2006 NYSESLAT administration by New York State teachers and experts, with guidance and assistance from the Department of Education and Harcourt Assessment. The New York Department of Education and Harcourt Assessment, Inc. conducted a fall field test administration of the updated NYSESLAT in October 2006; this information will be used to determine the reliability and validity of this test. The NYSESLAT is scheduled to be administered again in 2007.

Alignment to State Standards

The NYSESLAT is aligned to the state's English language Arts standards and the state's approved English as a Second Language (ESL) learning standards. Item mapping by New York State's English language learning standards by grade and domain are provided as evidence of test alignment.

In order to establish performance standards for the 2005 test administration, standard setting was conducted in spring 2005 in Albany, New York. Harcourt Assessment and the New York Department of Education led New York State–certified ESL, English language arts, bilingual education, and bilingual special education teachers through the standard-setting process. The item-mapping procedure was used to determine recommended cuts. This assessment is vertically scaled. Specific information on scoring and standard setting is located in the technical report.

Technical Properties of the Test

Item analysis. Item-level analyses for the 2005 administration of this test includes calculation of statistics based on classical test theory (CTT) such as *p*-values and point biserial correlations. In addition, item response theory (IRT) approaches, specifically Rasch model and partial-credit model statistics, were used in examining items. OUTFIT and INFIT, average Rasch difficulty by grade span by domain and standard error of Rasch difficulty values are provided in the technical manual. Specific information on item analysis can be found in the technical manual.

Test reliability. The reliability of the 2005 NYSESLAT was examined using classical test theory (CTT) and item response theory (IRT). Specifically, the use of Cronbachs' coefficient alpha statistic as well as the classical standard error of measurement (*SEM*) was applied to investigate the reliability of the test. Coefficient alpha reliability estimates were provided by grade and domain. Reliability ranged from a low of .64 in the kindergarten reading section to a high of .95 in grade 6 listening and speaking and grade 8 speaking. The conditional *SEM* values based on item response theory (IRT) are available in the technical manual. Rater agreement between both local raters and Harcourt raters was calculated for the pre-writing and writing constructed response items. Percent of agreement between the two different groups of raters, percentage difference scores between raters, means and standard deviations, the weighted kappa, asymptotic standard error, lower and upper 95% confidence limits, and the intraclass correlation were provided as measures of rater agreement. From second through the eleventh grade, the intra-class correlations are above .50; however, in the twelfth grade, the intra-class correlations range from 0.38 to .56 for the various reading response items. Reliability of classification of decision at proficient cut was also examined. Specific information on reliability is located in the technical manual.

Test validity. Validity for the 2005 NYSESLAT was reviewed in various ways. Content validity was established in the item development and review process. Item writers were trained to develop items representative of the standards embodied in the test blueprint. Items were also reviewed to ensure a match to instructional standards. In addition, item mapping also provided evidence of the match between standards and test items. The internal structure of the test was examined by calculating the biserial correlation coefficients and then examining the test items and the test blueprint to ensure that the appropriate constructs were being assessed. In addition, intercorrelations among the four domains by each grade were used as a measure of the test's internal structure. Please see the SELP summary for further information. To provide further validity evidence, Harcourt intends to investigate the relationship between the NYSESLAT and New York's English Language Arts Test for the 2006 administration. Specific information on validity is provided in the technical manual.

Freedom from bias. Items on the 2005 assessment were reviewed for bias by assessment specialists at Harcourt for freedom from bias. Specific information on freedom from bias is available in the technical manual.

RETURN to TABLE of CONTENTS

Technical Report

Harcourt Assessment (August 2006). *New York State Testing Program: NYSESLAT Technical Report.* Harcourt Assessment, Inc.

Ohio Test of English Language Acquisition (OTELA)

Grade Cluster(s): K; 1–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Ohio (spring 2006)

Test Purpose

The OTELA is used to determine English language proficiency of ELL students for Title III purposes.

Score Reporting

Scoring for the OTELA is similar to the ELDA. However, the OTELA does differ from the ELDA in that a revised cut score at the grade 3–5 cluster in writing was determined using a linear regression approach.

Test Development

Following the first operational administration of the English Language Development Assessment (ELDA), the Ohio Department of Education (ODE) contracted with the American Institutes for Research (AIR) to create two reduced-length ELDA forms per subject and grade cluster. The OTELA uses test items and scales from the ELDA, but has fewer test items within each domain. (For more information on the ELDA, see the summary for ELDA in this chapter.) The test was shortened by eliminating the easiest items from the ELDA. According to the ODE, the OTELA addresses the same English language proficiency standards as the ELDA and is comparable to the ELDA in terms of reliability. The ODE estimates that the time to administer the total test will be less than 40% of the time required to administer the ELDA. The OTELA was first fully implemented in Ohio in spring 2006.

Alignment to State Standards

According to the ODE, the OTELA is aligned to state English language proficiency standards. An alignment study could not be located.

Technical Properties of the Test

Item analysis. Rasch difficulty parameter estimates were calculated to assist in determining item difficulty. Average item difficulty estimates for the OTELA forms are very close to the targeted difficulties for the reading, listening and writing tests. Average item difficulties are provided in the technical manual.

Test reliability. Test developers first calculated the reliability of the test forms using the Spearman Brown prophesy formula. The reliability estimates ranged from .76 to .91 in test forms from grades 3–12. Next, they estimated the percentage of students at each test performance level. Finally, estimates were calculated for the classification consistency at each of the performance-standard cut scores as projected from 2005 ELDA operational test administration data.

Test validity. Additional validity studies could not be located for the OTELA.

Freedom from bias. Additional analysis to determine freedom from bias could not be located for the OTELA.

Technical Report

American Institutes for Research (May 2006). *The Ohio test of English language acquisition technical manual.*

Oregon English Language Proficiency Assessment (ELPA)

Grade Cluster(s): K–1; 2–3; 4–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Oregon (spring 2006)

Test Purpose

Oregon is using the ELPA to review the progress of ELLs for Title I and III accountability purposes. The ELPA was not designed to be administered for diagnostic or placement purposes.

Score Reporting

The ELPA is a web-based, computer-administered test that adjusts to each student's general level of proficiency. Embedded in the test is a locator phase to determine each student's general level of proficiency. After proficiency is determined, the remaining items are administered at

RETURN to TABLE of CONTENTS

the appropriate level for each student at the Beginning, Intermediate or Advanced level. Scores are provided in reading, writing, listening and speaking. In addition a comprehension score is derived from combining scores from listening and reading. An overall composite score is provided as well as a proficiency-level ranging from level 1 (defined as speaking little or no English) to level 5 (defined as full English proficiency).

Test Development

The State of Oregon was involved with the ELDA and Mountain West Assessment Consortium to develop their English language proficiency test. However, the Oregon Department of Education (ODE) decided against using items from either of these tests because they were not sufficiently aligned to Oregon's ELP standards. After this decision, the ODE began the process of developing the Oregon ELPA with Language Learning Solutions (LLS). Items were developed and piloted in May to June 2005. The first ELPA field test occurred from November 1 to December 2, 2005. From January 18 to February 3, 2006, a second field test took place. The first full assessment of all K–12 ELL students in Oregon was administered from April 4 to June 9, 2006.

During the ELPA's standard-setting process, teachers expressed some concerns with the test for students in kindergarten. As a result of these concerns, the ODE is considering developing a separate kindergarten test. Teachers also expressed concerns about test alignment, about how the medium of testing might interfere with student performance, and about the need to address academic English. In light of these further concerns, in fall 2006 the ODE wrote new items which will be part of a pilot test in spring 2007. After this pilot test, standard setting will be undertaken again.

Alignment to State Standards

According to the ODE, the ELPA is aligned to Oregon's English language proficiency standards as well as state content standards. An alignment study could not be located.

Standard setting was conducted in summer 2006 by educators, parents, and community members using the bookmarking technique. Additional information on scoring and standard setting could not be located.

Technical Properties of the Test

Item analysis/test reliability/test validity/freedom from bias. Information on item analysis, test reliability, validity, and bias analyses could not be located.

Technical Reports

The technical manual is currently not available. The State of Oregon has contracted with the American Institutes for Research (AIR) to produce the technical report. It will be available in 2007.

The Stanford English Language Proficiency Test (SELP, Stanford ELP)

Grade Cluster(s): Pre–K; K–1; 1–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2003–2006
State(s) Using This Test for Title III Accountability (Implementation Date): Off-the-shelf version: Mississippi (2003–2004). Augmented versions: Arizona (AZELLA 2006); New Mexico (NMELPA, 2006); South Dakota (Dakota ELP, 2006); Virginia, (spring 2006); Wyoming (WELLA, spring 2006)

Test Purpose

The purpose of the Stanford ELP is to measure English language proficiency development in pre-K–grade 12 English language learners.

Score Reporting

Scores are provided in the following domains: listening, speaking, and reading, and total writing (a combined score based on the writing and writing conventions domains). Additionally, scores in the following skills are given, based on the combined results of the following domain scores: Productive (speaking and writing), Comprehension (listening and reading), Social (listening and speaking), and Academic (reading, writing, and writing conventions). A total composite scaled score is also provided. Based on test results, students are placed into one of the following five proficiency levels: Pre-Emergent, Emergent, Basic, Intermediate, and Proficient, Pre-Emergent being the lowest performance level and proficient being the highest.

RETURN to TABLE of CONTENTS

Test Development

Harcourt Assessment conducted two standard-setting sessions to determine cut scores, the first taking place on May 21, 2005 for the early version of the complete Stanford ELP and SLP Preliteracy level. The second standard-setting panel convened on October 22, 2005 to review the Stanford ELP and SLP Readiness/Preescolar level and the revised Stanford ELP and SLP Preliteracy/Preprimario level. Content experts reviewed the assessment and established preliminary cut scores using the modified Angoff procedure. The SELP is a vertically scaled assessment.

Harcourt Assessment Inc. first published the SELP in 2003. Two Stanford ELP test forms, Forms A and B, were published in 2003 in the primary, elementary, middle grades, and high school levels. Form C was later published in 2004 in these levels. The components of the Stanford ELP Readiness and Preliteracy levels, Form A, were published from 2004–2006. Some states are using the catalog version of the SELP while others are using augmented SELP tests. Augmented tests contain a core of catalog SELP items and additional test items written to the state's specifications in each domain. Test developers use Rasch or partial-credit models to link the augmented test to the Stanford ELP scale through the use of actual examinee data. Delaware and Mississippi are using the catalog SELP; Wyoming is using the SELP Form A in grades K–2 only. States using augmented versions of the SELP include: 1) Arizona: Arizona English Language Learner Assessment (AZELLA) 2) New Mexico: New Mexico English Language Assessment (NMELA) 3) South Dakota: Dakota ELP 4) Virginia: augmented SELP and catalog Form C 5) Washington: WLPT-II and 6) Wyoming: Wyoming English Language Learner Assessment (WELLA) in grades 3–12 only.

Alignment to State Standards

The Stanford ELP test is based on the Teachers of English to Speakers of Other Languages (TESOL) standards (1997) as well as individual state English as a Second Language (ESL) standards. Delaware, which uses an off-the-shelf version of the SELP, hired an outside researcher to conduct an alignment study between the SELP and the state-adopted English language arts content standards and Grade-Level-Expectations. This alignment study, which used the Webb Alignment Tool, concluded that the SELP was not sufficiently aligned to Delaware's ELA content standards and Grade Level Expectations. Other alignment studies have or are being conducted by states that are using the SELP. Many of these alignment studies have also shown the need for augmented version of the SELP in order to align with their own state's standards. Please see the summaries on these customized state assessments for more information about alignment with ELA standards and state content standards.

Technical Properties of the Test

Item analysis. Item analyses determined item discrimination and item difficulty. *P*-values were provided as a measure of test difficulty. For Primary Form A-Listening, the average *p*-value was .84. At the Primary level for Form B-Listening, the average *p*-value was .81. At the Primary level for Form C-Listening, the average *p*-value was .78.

Test reliability. Indices of internal consistency and alternate-forms reliability, as well as standard errors of measurement values, were calculated for SELP during the field testing of items and the tryout of forms. For Forms A and B, the alpha of the whole test at the Primary level was .94, at the elementary level .94, for Middle grades .94, and for high school .93. For Form C, the alpha of the whole test at the Primary level was .93, in Elementary .92, in the Middle grades .94, and High School .96. Additional information on reliability was not available.

Test validity. Content validity was established through expert review. Item writers were trained to write items aligned with instructional standards set forth in the test blueprint. In addition, the review process included examining the match between the item and the relevant instructional standard. Items relating specifically to instructional standards were included in the test forms. Criterion-related validity was established by correlating scores on the SELP to scores on other tests. The Stanford ELP was correlated to the Stanford Diagnostic Reading Test (SDRT) and Stanford Achievement Test Series, Ninth Edition (SAT 9). There was a strong correlation between scores earned on the Stanford ELP and the SDRT. There was a low positive correlation between scores earned on the Stanford ELP multiple-choice domains and the Abbreviated Stanford 9 reading domain. Construct validation was established through the examination of correlation coefficients among SELP subscore combinations. Since the publication of the original technical manual, the Stanford ELP internal structure, a validity study using factor analysis has been conducted. Results which supported the theorized test structure will be detailed in the 2007 revision of the technical manual.

RETURN to TABLE of CONTENTS

Freedom from bias was addressed through formal review of the SELP by experts of cultural and linguistic bias. Harcourt held a bias and sensitivity review of the test with an advisory board of ESL experts. All SELP test items were reviewed after they were assembled into field test forms. The panel suggested changes to items, directions, and format. Following the advisory board's review for bias, further revisions were made to the items as needed. DIF analysis was not conducted due to limitations in sample size. Additional information on freedom from bias was not available.

Technical Reports

The technical report for the SELP will be available from Harcourt Assessment Inc. in 2007.

Test of Emerging Academic English (TEAE)

Grade Cluster(s): 3–4; 5–6; 7–8, 9–12
Domains Tested: Reading and writing
Date(s) Published: 1998; 2004
State(s) Using This Test for Title III Accountability (Implementation Date): Minnesota (1998)

Test Purpose

The TEAE is designed to evaluate students' progress towards English proficiency, regardless of the programs they are enrolled in and the classroom settings in which they are taught. It was revised in 2004–2005 for the added purpose of meeting Minnesota's Title III requirements under NCLB.

Score Reporting

Separate raw scores, scale scores and proficiency levels (1–4 for reading and 1–5 for writing) are reported for each domain.

Test Development

TEAE was first developed in 2000 to address Minnesota's statewide mandate to measure progress in English reading and reading proficiency for students whose first language is not English. The TEAE, developed by the Minnesota Department of Education (MDE) and MetriTech Inc., was based upon the Language Proficiency Test Series. Adapted by the test developer in consultation with Minnesota teachers, TEAE was first administered in fall 2001.

Standard Setting

A large expert panel met initially in 2002 to determine cut scores for the TEAE. This panel used IRT-based item maps (taken from the 2001 test administration) as part of a modified bookmarking procedure to set initial cut scores for the TEAE. These initial cut scores were adjusted in a second meeting after a smaller committee of general education and ESL teachers developed a clearer set of performance-level descriptors for the larger standards-setting panel. Another standard-setting study was conducted in winter 2003 to review preliminary cut-scores set in 2002. This time, the panel used item maps, ordered item booklets, and performance-level descriptors to set cut scores for the TEAE domains. The four test forms (for grades 3–4, 5–6, 7–8, and 9–12) are vertically scaled, while scores from different grades are placed on a common scale.

In June 2003, Minnesota developed and adopted its own English Language Proficiency Standards to meet the federal requirements of the No Child Left Behind Act. A newer version of the TEAE, TEAE-II, is being developed to align closely with Minnesota's content standards, with the Minnesota English Language Proficiency (ELP) Standards for English Language Learners K-12, and to meet Title I and Title III requirements.

Alignment to State Standards

A study was conducted by the Minnesota Department of Education to determine the degree of alignment between Minnesota's ELP standards and the TEAE (Minnesota Department of Education, 2004). A panel of five ESL professionals rated all 500 test items that were developed for all forms and grade-level cluster tests. Using a standardized protocol, each item was rated for its alignment with the four language components described by Minnesota's ELP standards: 1) purpose, audience & genre; 2) communicative functions; 3) language features; and 4) word knowledge & use. Evaluation of each test item also included experts' evaluation of which proficiency level was indicated for each item at each grade-level cluster. Alignment of items by grade cluster showed that the test items were loaded at the intermediate, advanced, and transitional proficiency levels. MDE staff concluded from this study that the TEAE is suited for Minnesota's English learner population, since three-quarters of these students are at the intermediate level of proficiency or above.

RETURN to TABLE of CONTENTS

In 2003, the MDE conducted an alignment study of the TEAE and Minnesota's grade-level expectations for reading for grades 3, 5, 7, and 10. Using Webb's alignment procedure, the MDE found that the TEAE aligned with statewide reading assessments in the following areas: categorical concurrence, depth of knowledge, and depth of alignment.

Technical Properties of the Test

Item analysis. Item analysis could not be located.

Test reliability. Cronbach's alpha coefficients and the standard error of measurement (*SEM*) were reported for TEAE scores in the 2007 technical manual. Alpha coefficients were reported by domains, by grades, and by students' number of years in Minnesota schools. Coefficients for the total raw reading score ranged from .88 to .97. Reliability coefficients for overall reading scores appeared to go down slightly among students who had attended Minnesota schools for five years or more. Alpha coefficients for the overall raw score for the reading domain ranged from .77 to .96, with reliability coefficients decreasing slightly among students who had been in Minnesota schools for five years or more. Also reported were the inter-rater reliability statistics given by percent agreement and the Pearson product-moment correlation for the reading section. Please see the technical manual for further information on the reliability of this assessment.

Test validity. The MDE conducted two validity studies on the TEAE. The first consequential validity study was conducted in summer 2002 for reading and in winter 2003 for the reading domain. This study focused on the test performances of those students from the two largest districts in the state who placed in the highest proficiency levels on the TEAE. The students' scores on the TEAE were compared to their test scores, their classroom performance, and other state measures. Confounding factors, such as lack of motivation on the part of secondary students, were discussed in the findings of the report.

In the second consequential validity study, district administrators representing the two largest urban districts, in conjunction with a rural and suburban superintendent, reviewed the results of the standard-setting study described in the Test Development section, above. These administrators were presented with the item maps, descriptors, and correlation statistics that had been provided to the standard-setting panel. As a result of this consequential validity study, minor changes were made in the cut-scores for the proficiency levels for grade clusters in the reading and writing sections of the TEAE.

Albus et al (2004) conducted a criterion validity study comparing 99 ELL scores on the TEAE, the Minnesota Basic Skills Test (BST), and the Minnesota Comprehensive Assessments (MCA). This study found strong correlations between TEAE Reading and MCA Reading scores (.693), TEAE Reading and BST Reading (.701), and between TEAE Writing and BST Reading (.738). However, the authors cautioned that the TEAE Writing and BST Reading correlation was based upon a relatively small sample size.

Freedom from bias. The technical manual is unclear as to whether or not differential item functioning analysis was performed using the Mantel-Haenszel procedure by gender and ethnicity. Results of these analyses could not be located.

Technical Reports

The Minnesota Department of Education and Pearson Educational Measurement (2007, January). *The Minnesota assessments technical manual for the academic year 2005–2006*. St. Paul, MN: Author.

Minnesota Department of Education (2003, 2005). Minnesota English language proficiency standards for English language learners, K-12. Retrieved September 2, 2007 from http://education.state.mn.us/mdeprod/groups/EnglishLang/documents/Report/002201.pdf

Texas English Language Proficiency Assessment System (TELPAS)

Grade Cluster(s): Reading Proficiency Test in English (RPTE): 3; 4–5; 6–8; and 9–12. Texas Observation Protocols (TOP): each individual grade from K to 12

Domains Tested: Reading Proficiency Test in English (RPTE): 3-12 reading. Texas Observation Protocols (TOP): K-2 reading; K-12 speaking, listening, and writing

Date(s) Published: Reading Proficiency Test in English (RPTE): 2000. Texas Observation Protocols (TOP): 2005

State(s) Using This Test for Title III Accountability (Implementation Date): Texas (2005)

Test Purpose

The TELPAS system consists of two components: the Reading Proficiency Test in English (RPTE) and the Texas

RETURN to TABLE of CONTENTS

Observation Protocols (TOP). Since 2005, the TELPAS results have been used in the Annual Measurable Achievement Objective (AMAO) accountability measures required by NCLB. The TELPAS is used to measure students' annual progress in learning English in listening, speaking, reading, and writing. The TELPAS system is also used in combination with other measures to make instructional decisions for individual students. Beginning the 2007-2008 school year, only the TELPAS acronym will be used for RPTE and TOP.

Score Reporting

The four TELPAS proficiency ratings are: Beginning, Intermediate, Advanced, and Advanced High. Students are given a proficiency rating in reading, writing, listening and speaking. A comprehension rating is also given; the listening and reading ratings are each converted to a number from 1 (Beginning) to 4 (Advanced High). The average of the two numbers is the comprehension score. An overall composite level of proficiency, which combines the results of all four language domains, is also given. The language domain of reading is given most weight in the composite rating, followed by writing, listening and speaking have the least weight. The composite score ranges from 1 (ratings of Beginning in all language areas) to 4 (ratings of Advanced High in all language areas).

TOP is holistically scored; skills are not assessed in isolation. The TOP Proficiency Level Descriptors are the holistic scoring rubrics used by teachers to give one of four proficiency ratings in each of the four domains of reading, writing, listening and speaking for K-2 and listening, speaking and writing for grades 3-12.

Test Development

The RPTE was originally developed in response to Texas state regulations passed in 1995. Based on the recommendations of an advisory committee of assessment specialists and content experts, the Texas Education Agency (TEA) developed prototype test items in conjunction with Pearson Educational Measurement and Beck Evaluation and Testing Associates (BETA), the test contractors. The resulting items were field tested in the spring of 1999. In the fall of 1999, TEA conducted a field study to determine the test format and length. Following the spring 2000 test administration, raw score ranges for each proficiency level were established by TEA in conjunction with external assessment and content experts and practitioners based on second language acquisition theory and statistical analyses of student performance. Scaling of the assessment was conducted in fall 2000.

New items are written each year and reviewed by educators in the State of Texas. These items are then field tested in spring of each year. The TEA has undertaken the development of a second edition of RPTE beginning in the 2004–2005 school year. This second edition will add a second-grade assessment form and change the grade clusters to 2, 3, 4–5, 6–7, 8–9, and 10–12. This revised version will assess more of the type of reading required in the subject areas of science and mathematics. Field-testing of the second edition took place in spring 2006 and 2007, and the new edition will be implemented in spring 2008.

The Texas Education Agency (TEA), in conjunction with its testing contractor Pearson Educational Measurement, developed the TOP to assess the federally required domains and grade levels not tested on the RPTE. TOP was created by TEA along with test development contractors, bilingual/ESL consultants, and members of an English language learner focus group composed of teachers, bilingual/ESL directors, assessment directors, campus administrators, and university professors. TOP assesses students through observations in an authentic environment as students engage in regular classroom activities. In grades 2–12, the writing domain is assessed through a collection of classroom-based writing. The test was benchmarked in 2004 and fully implemented beginning in 2005.

Standard Setting

The TEA and its testing contractors, technical experts and second language acquisition experts, an English language learner (ELL) assessment focus group of Texas educators and administrators from regional, district, and campus levels, and other Texas professional educators assisted in creating composite rating weighting formulas for the 2005 and 2006 TELPAS assessments to determine cut scores for each of the four proficiency levels within each domain and for the overall proficiency ratings. Additional information on scoring and standard setting is available in the technical report.

Alignment to State Standards

The RPTE was developed to align with the state's previous assessment program, the Texas Assessment of Academic Skills (TAAS). Beginning in spring 2004, RPTE was revalidated to be more closely aligned with the Texas Assessment of Knowledge and Skills (TAKS) reading selections and test

RETURN to TABLE of CONTENTS

questions. The TAKS, in turn, was developed to align with state content standards, providing a link between RPTE and Texas' content standards. The Texas Education Agency reports that the RPTE II, which will be fully implemented in 2008, will be aligned to the Texas content standards for reading and the English language proficiency standards, which emphasizes academic English acquisition.

Technical Properties of the Test

Item analysis. Each RPTE test question and reading selection is developed to align with proficiency level descriptors that are the foundation for test development and test construction. Before and after field testing, committees of educators review the reading selections and items to eliminate potential bias and ensure appropriateness in terms of content, age appropriateness, and proficiency level alignment. To determine the quality of each test item, the testing contractor produces statistical analyses for each using 3 types of differential item analyses: calibrated Rasch difficulty comparisons, Mantel-Haenszel Alpha and associated chi-square significance, and response distributions. Point biserial data are also evaluated yearly for each test item. Additionally, in order to ensure that the items perform consistent with the proficiency level descriptors and discriminate between students in the various proficiency level categories, the p-values of students in each proficiency level category are examined for each field-tested item. The educator review committees are provided with these data for each item that is field-tested in the annual field-test items review procedure. Using this information, item review committees review newly developed items for appropriateness of each item for use on future tests.

Test reliability. Internal consistency, the standard error of measurement (SEM) and the conditional SEM were calculated. Reliability estimates were also reported for items from the 2005–2006 test administration. Reliability is expressed in stratified alpha reliability for tests/objectives involving short-answer and/or essay questions; KR-20 reliability was computed for all other question types. These reliabilities are provided by grade and by grade/gender. Reliability coefficients are reported for grades 3, 4-5, 6-8, and 9-12.

A large-scale study of rating validity and reliability of the TOP was conducted by TEA in spring of 2006. An audit of more than 13,000 scored writing samples collected from teachers who were trained TOP raters was conducted to evaluate how effectively raters applied the rubrics. Individuals trained as TOP raters at the state level rescored the student writing collections. Overall the state and teacher ratings agreed perfectly 77% of the time. The study also required the raters of the students selected for the audit to complete a questionnaire concerning the adequacy of the training and scoring processes for each language domain. Of the more than 6,000 raters audited, following are the percents of raters indicating that the training provided them with sufficient information to judge the English language proficiency levels of their students in each language domain: listening 96%, speaking 96%, writing 97%, and reading (grade 2 only) 94%. Detailed information on this study is available in the technical report.

Test validity. Two studies examined the relationship between RPTE and TAKS performance levels. The first study, which took place after the spring 2004 test administration examined the following issues: 1) the percent of qualifying recent immigrants who met the AYP incremental progress performance standard in spring 2004, and 2) the reading performance of LEP students evaluated under the incremental progress model compared to that of LEP students evaluated with TAKS and 3) the instructional rationale for incremental RPTE progress model. These statistical alignment analyses indicated that the percentages of students who met the RPTE incremental progress standard and TAKS standard were very similar. A second study undertaken after the spring 2005 test administration established a connection between RPTE scores and the TAKS performance categories of Met Standard (passing level) and Commended Performance (highest performance level). In addition, content validation studies are conducted yearly by panels of Texas teachers, test development specialists and TEA staff members. Specific information on test validity is given in the technical digest.

Freedom from bias. Please see technical manual for details of how freedom from bias issues were addressed for each test in the TELPAS system.

Technical Reports

Technical Digest 2004–2005. *Student assessment division*. Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Technical Digest Texas English Language Proficiency Assessment System (TELPAS) 2004–2005. Appendix 7: *Development of the TELPAS composite ratings and composite scores*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/ListofAppendices/TechDigest-A7.pdf

Texas Assessment. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Education Agency (2002). *Student Assessment Division: Technical Digest 2001–2002*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig02/index.html

Texas Education Agency (2003). *Student Assessment Division: Technical Digest 2002–2003*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig/index.html

Texas Education Agency (2004). *Student Assessment Division: Technical Digest 2003–2004*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig04/index.html

Texas Education Agency (2005). *Student Assessment Division: Technical Digest 2004–2005*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Texas Assessment (2006a). *Student assessment division: Technical digest 2005–2006*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Assessment (2006b). Student Assessment Division: Technical Digest 2005–2006. *Appendix 6*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Assessment (2006c). Student Assessment Division: Technical Digest 2005–2006. *Chapter 15: Validity*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/Chapters/Chapter15_Validity.pdf

Texas Assessment (2006d). Student Assessment Division: Technical Digest 2005–2006. *Chapter 17: Reliability*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/Chapters/Chapter14_Reliability.pdf

Texas Assessment (2006e). Student Assessment Division: Technical Digest 2005–2006. *Appendix 10*. Retrieved September 3, 2007 from: http://k12testing.tx.ncspearson.com/TechDigest/ListofAppendices/TechDigest-A10.pdf

Technical Digest 2004–2005. *Student assessment division*. Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Technical Digest 2004–2005. Appendix 7: *Development of the TELPAS composite ratings and composite scores*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/ListofAppendices/TechDigest-A7.pdf

Texas Assessment. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Utah Academic Language Proficiency Assessment (UALPA)

Grade Cluster(s): K; 1–2; 3–6; 7–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Utah (fall 2006)

Test Purpose

The purpose of the UALPA is to provide a total scaled proficiency score to educators, schools, districts, and states and for Title III reporting purposes.

Scoring Reporting

A raw score is provided in reading, writing, listening and speaking. A Comprehension raw score is also calculated. This score is a composite of reading and listening scores. A composite score is provided to determine overall proficiency. During the 2006–2007 school year, students will be placed in the following proficiency levels, which are based on Idea Proficiency Test (IPT) descriptors: A, B, C, D, and E. A linking study will provide the correlation between the IPT and the UALPA. At the end of 2006-2007 and a standard setting, students will be identified using the UALPA proficiency levels of, P – Pre-Emergent, E – Emergent, I – Intermediate, A – Advanced, and F – Fluent. With the beginning of the 2007-2008 school, students will be identified by the UALPA proficiency levels and will be assessed once a year for growth using the UALPA.

Test Development

The UALPA was created in collaboration with Measured Progress and the Mountain West Assessment Consortium. For more information on initial test development please see the summary for the Mountain West Assessment Consortium in this chapter. The UALPA was administered in the State of Utah from October 30, 2006 to February 28, 2007 to 50% of English learner students in the state. The

RETURN to TABLE of CONTENTS

remaining 50% of students will be tested from March to April 30, 2007. Students who moved into school districts after April 30th were also assessed using the UALPA. Their scores were not included in either the linking study or standard setting.

Formal proficiency-level cut scores were not determined for the MWAC English proficiency assessments. Measured Progress led a group of national experts in English language acquisition in recommending cut scores for state panels. A modified bookmark method for standard setting was used to recommend cut scores. Ultimately, however, it was recommended that states conduct standard setting using their own data.

The UALPA was pulled from implementation in the spring of 2006 to refine the development of the test. At the conclusion of the year a standard setting was held by the USOE in conjunction with Questar Assessment Inc. Previous to the plan was presented to the state's TAC committee. Revisions were suggested and adopted. Data analysis and state standards-adoption processes were parallel for all grades and levels of the assessment

A standard setting was held June 26-28, 2007. Two panels were convened to recommend standards for UALPA using the book mark method. One panel recommended standards for grades K through 6; the other panel made comparable recommendations for grades 7 through 12. The 35 standard setting participants represented a broad range of stakeholders – both active educators and non-educators. Panel members included classroom teachers (both English Language Learner instructors and general education), building and district administrators, parents, curriculum directors, related professional services staff, and other representatives of the general public.

The methodology used for all sessions was "item mapping" or "Bookmark Procedure". Final cut scores were determined by using item maps and a "Bookmarking Procedure".

After the completion of the panel sessions a conference was held with Questar and USOE to review impact data. The USOE reviewed the panels' recommendation and the impact data. Levels were adjusted in order to improve the consistency of outcomes across grades and to maintain the appropriate amount of rigor associated with a designation of 'Advanced'. UALPA-1 results will be reported as Pre-emergent, Emergent, Intermediate, or Advanced. The test, as currently configured, does not provide sufficient evidence to classify performance as 'Fluent'.

Alignment to State Standards

State standards were revised during the 2006 -2007 year. The standards have gone through a number of committee reviews, content expert and ELL educators' reviews. A holistic alignment study was completed in June 2007 by WestEd. The alignment study revealed a strong correlation to grade level skills. However, some of the standards did not reflect an appropriate progression of English Language development skills across proficiency levels. The standards will continue to go through refinement during the 2007-2008 year.

Technical Properties of the Test

Item analysis. Please see the summary on the Mountain West Assessment in this chapter for information on initial item analysis. Information on subsequent item analyses could not be located.

Test reliability/test validity. Information on test reliability and validity could not be located.

Freedom from bias. Please see the summary for the Mountain West Assessment in this chapter for more information on initial freedom from bias analyses. Information on subsequent freedom from bias analyses could not be located.

Technical Reports

The technical manual is currently not available. The next test contractor, Questar, is expected to complete the technical manual in 2008.

Washington Language Proficiency Test II (WLPT-II)

Grade Cluster(s): K–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Washington (2006)

Test Purpose

The State of Washington uses the WLPT-II to determine growth of ELLs for Title III reporting purposes as well as to determine if a student is ready to exit the English language program.

RETURN to TABLE of CONTENTS

Score Reporting

Scale scores are provided in reading, writing, listening and speaking. A composite score is also calculated; this score is a combination of scores across domains. Based on test results, students are placed in one of four proficiency levels: Beginning/Advanced Beginning, Intermediate, Advanced, and Transitional.

Test Development Summary

The WLPT-II is an augmented and aligned version of the Harcourt Stanford English Language Proficiency (SELP) test. The WLPT-II consists of both SELP Form A items and augmented items created by Washington teachers in October 2005. Augmented items were field tested in 2006. Based on the performance of each item, items were included or dropped in the 2006 operational form. The 2007 WLPT-II Form B will be created using items from SELP Form B, and the 2008 WLPT-II Form C will be created using items from SELP Form C.

The bookmarking technique was used in the standard-setting process. Additional information on scoring and standard setting is available in the technical manual

Alignment to State Standards

The WLPT-II is aligned to Washington State English language development standards. Harcourt conducted an alignment study using item mapping to compare the SELP forms to the state's ELD standards. In September 2005, a second alignment study was conducted by Washington State educators using the state's English language proficiency descriptors, to identify general gaps in the SELP forms. This committee recommended the augmentation of the SELP forms in reading, writing, and speaking so that the test would be aligned to Washington State ELD standards. Approximately 20% of core items were revised or replaced on the test. Additional information on test alignment is provided in the technical manual.

Technical Properties of the Test

Item analysis. Items were assessed using a classical test theory (CTT) and item response theory (IRT) framework. CTT statistics provided include *p*-values calculated to detect item difficulty, and point-biserial correlations were used to determine test item discrimination. Average outcomes for these analyses could not be located. IRT approaches include calculation of Rasch difficulty, INFIT and OUTFIT. Additional information on IRT statistics is provided in the technical manual.

Test reliability. The reliability of this assessment is shown by Cronbach's alpha, the classical standard error of measurement (*SEM*), and the conditional *SEM* from item response theory. Reliability statistics were provided in each of the four domains. For listening, reliability across grade spans ranged from 0.69 to 0.81; in reading, reliability ranged from 0.75 and 0.86 across grade spans. For speaking, the reliability ranged from 0.91 to 0.96, and for reading, it varied from 0.75 to 0.86. Overall, the speaking test was the most reliable across grades. The mean reliability for the test overall was .94, with a range of reliability from .89 to .95. Intrarater and inter-rater agreement was also assessed. Targeted agreement for calibration responses was met for intrarater reliability; at least 80 percent were in perfect agreement, plus 20 percent adjacent agreement was also achieved. The targeted agreement rate for calibration responses for inter-rater reliability was 70% perfect agreement with no more than 5% of greater than +/- 1 score point discrepancy; this goal was exceeded. Additional information on reliability is available in the technical manual.

Test validity. The validity of this test was determined in several ways. Content validity of SELP and augmented items were reviewed by Harcourt experts, ESL experts, the Washington Department of Education and Washington ESL professionals to ensure that all items aligned to the state's English language development standards. To assess the test's construct validity, intercorrelations between domains were calculated. Point biserial correlation coefficient and fit statistics are also offered as evidence of test validity. Unidimensionality of the assessment was determined through principal components analysis for each grade scan. The analysis verified the unidimensionality of the assessment. These statistics and additional validity information are available in the technical manual.

Freedom from bias. A committee composed of Washington State ESL experts reviewed the SELP forms in August 2005 to ensure that the test was free from bias. Based on this review, the committee recommended the revision of some items. State educators also reviewed the new augmented items for bias and fairness. Differential item functioning (DIF) was assessed by gender using the Mantel (1963) extension of the Mantel-Haenszel procedure for the open-ended items and the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) for the multiple-choice items. Open-ended items were analyzed using the Mantel statistic with the standardized mean difference

RETURN to TABLE of CONTENTS

(SMD). Additional information on freedom from bias is available in the technical manual.

Technical Reports

Washington Department of Education (Dec. 2006). *Washington language proficiency test - II* (Technical Report 2005–2006 School Year). DRAFT. Olympia, WA: Harcourt Assessment.

West Virginia Test for English Language Learning (WESTELL)

Grade Cluster(s): K–2; 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening, and speaking
Date(s) Published: 2005
State(s) Using This Test for Title III Accountability (Implementation Date): West Virginia (2005)

Test Purpose

Please see the ELDA section in this chapter for test purpose information.

Score Reporting

Please see the ELDA section in this chapter for information on scoring and standard setting.

Test Development Summary

The state of West Virginia currently uses the English Language Development Assessment (ELDA). However, in 2005 the state Board of Education adopted policy language that names the state's ELP assessment as the West Virginia Test of English Language Learning (WESTELL). Please see the ELDA section in this chapter for test development information, including item development and standard setting.

Alignment to State Standards

Please see the ELDA section in this chapter for information on alignment to state standards.

Technical Properties of the Test

Please see the ELDA section in this chapter for information on item analysis, test reliability and validity, and freedom from bias studies.

Technical Report

Please see the ELDA section in this chapter for information on the technical report.

Wyoming English Language Learner Assessment (WELLA)

Grade Cluster(s): 3–5; 6–8; 9–12
Domains Tested: Reading, writing, listening and speaking
Date(s) Published: 2006
State(s) Using This Test for Title III Accountability (Implementation Date): Wyoming (spring 2006)

Test Purpose

The purpose of this assessment is to measure annual growth of English language learners in meeting English language proficiency standards. In addition, the assessment is used for student identification and placement in the fall. WELLA is used diagnostically, helping teachers to focus their teaching to specific areas of student need.

Score Reporting

See the summary on the SELP in this chapter for more information on the scoring of this assessment.

Test Development Summary

The WELLA is an augmented version of the Stanford English Language Proficiency test (SELP). The test consists of existing SELP items and additional items that were specifically created for the State of Wyoming by Harcourt Assessment. These newly developed items include math and science items produced by Harcourt Assessment content experts. The WELLA was field tested in Wyoming in spring 2005 and was fully implemented in the Wyoming in spring 2006.

Alignment to State Standards

Harcourt Assessment Inc. evaluated the alignment between the SELP and Wyoming ELD standards and determined there is a strong match. However, some standards could not be easily assessed through a large-scale paper-and-pencil test so these standards were removed. Information was not available about when the alignment study was conducted and what procedures were used to conduct the study. The Wyoming Department of Education is currently conducting an independent alignment study of the WELLA to determine how the test aligns to state ELD standards.

Technical Properties of the Test

Item analysis. P-values were provided as a measure of item difficulty. Please see the technical report for more information on item analysis for the WELLA.

RETURN to TABLE of CONTENTS

Test reliability/test validity. Information on the reliability and validity of this test could not be located

Technical Reports

Harcourt (2007). WELLA technical manual (DRAFT). pp. 1–6. The technical manual on the WELLA is being written by Harcourt Assessment, Inc. and will be available in 2007.

Conclusion

After tremendous effort in a short period of time, states and test developers have made progress in complying with NCLB Title III stipulations. It is important that the reliability and validity of these assessments be examined and the tests be refined even further. Assessments need to undergo rigorous psychometric analyses on an ongoing basis. Continued partnerships between test developers, states and researchers will raise the psychometric standards for English language proficiency tests. In addition, further efforts must be made to use alignment methods in the test development process to ensure that tests are valid and reliable measures of state English language development standards and state-adopted content standards.

The research team would like to thank those representatives of the various state departments of education and test developers for their time and cooperation in assisting us with gathering these data. This chapter would have been impossible to write without their help. For a complete list of the English proficiency assessments in this chapter, please see Table 1.

Additional Sources

No Child Left Behind Act of 2001 (NCLB). Pub. L. No. 107–110, § 115 Statute 1425. (2001).

Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS):

ACCESS for ELLs®: 2005–2006 Informational handbook for Wisconsin district assessment coordinators and bilingual/ESL administrators. Retrieved September 11, 2006, from http://dpi.wi.gov/oea/doc/ells_ACCESSinfo_hdbk.doc

Accommodations for ACCESS to ELLs. Retrieved September 10, 2006, from http://www.wida.us/ACCESSForELLs/accommodations/view?searchterm=

Alignment of model performance indicators and versatility of frameworks. Retrieved September 10, 2006, from http://www.wida.us/Resources/ELP_Standards_Overview/section_03.html

Cook, G. (2006). *Findings of an alignment study of the Stanford English language proficiency test and the English language proficiency (World-class instructional design and assessment) to the Delaware grade level expectations in English language arts for grades 3, 4, 5, 6, 7, 8, 9, and 10. Submitted to the assessment and accountability branch of the Delaware Department of Education.* Retrieved July 5, 2007, from http://www.doe.k12.de.us/aab/DE%20SELP-ELP%20Alignment%20Report--Final.pdf.

Davidson, F., Kim, J.T., Hyeong-Jong, L., & Li, J. (2006). *New Jersey alignment study*. WIDA Consortium. *Frequently asked questions regarding ACCESS for ELLs®* Retrieved August 20, 2006, from http://www.maine.gov/education/esl/AccommodationsforACESSforELLs.htm

Gottlieb, M. et al (2007). *ACCESS for ELLs® Interpretive guide for score reports (Spring 2007).* Madison: Board of Regents of the University of Wisconsin System–University of Wisconsin Center of Education Research.

Gottlieb, M. (2006). *Interpretive guide for score reports*. Madison, WI: World-Class Instructional Design and Assessment Consortium and Wisconsin Center for Educational Research.

Understanding the ACCESS for ELLs® test. Retrieved August 20, 2006, from http://www.wida.us/ACCESSForELLs/

Arizona English Language Learner Assessment (AZELLA):

Arizona Department of Education (2006a). *AZELLA summer 2006 trainings power point presentation.* Retrieved November 20, 2006, from http://www.ade.state.az.us/asd/lep/downloads/AZELLAWorkshopPresentation.pp

RETURN to TABLE of CONTENTS

Arizona Department of Education (2006b). *AZELLA summer 2006 trainings sample items power point presentation.* Retrieved November 20, 2006, from http://www.ade.state.az.us/asd/lep/SampleItems.ppt

California English Language Development Test (CELDT):

California Department of Education (2007, March). *Assistance packet for school districts and schools.* Retrieved March 29, 2007, from http://www.cde.ca.gov/ta/tg/el/documents/celdt07astpkt.pdf

Katz, A., Low, P., Stack, J., & Tsang, S. (2004). *A study of content area assessment for English language learners*. Prepared for the Office of English Language Acquisition and Academic Achievement for Limited English Proficient Students, U.S. Department of Education. ARC Associates, Inc: Oakland, CA. Retrieved September 5, 2006, from http://www.arcassociates.org/files/CAELLRpt9-04.pdf

California Department of Education (2006). *2006 Score Reports and Interpretation Guides*. Retrieved March 29, 2007, from http://www.cde.ca.gov/ta/tg/el/resources.asp

California Department of Education (2006, July). *CELDT Initial/annual scale ranges*. Retrieved March 29, 2007, from http://www.cde.ca.gov/ta/tg/el/cutpoints.asp

California Department of Education (2006a, March). *California English language development test: Performance level standard setting last minute memorandum.* Retrieved March 29, 2007, from http://www.cde.ca.gov/be/ag/ag/yr06/documents/bluemar06item11.doc

California Department of Education (2006b, March). *California English language development test: Performance level standard setting: State board of education March 2006 Item 11 (DOC)*. Retrieved March 29, 2007, from http://www.cde.ca.gov/be/ag/ag/yr06/documents/mar06item11.doc

Colorado English Language Assessment (CELA):

Colorado Board of Education (2006). *Action agenda item from the March 6, 2006 meeting*. Contract: Colorado Department of Education, CTB/McGraw Hill, Office of Management Services of the Colorado English

Colorado Department of Education (2007). *CELA Proficiency Test 2007*. PowerPoint presentation. Retrieved from, http://www.cde.state.co.us/cdeassess/documents/cela/index_cela.html.

Colorado English Language Assessment (CELA). Retrieved September 17, 2006, from http://gsbaeboard.org/cgi-bin/WebObjects/CDEAgenda.woa/wo/

Medina, B. (2006a). *Memorandum to superintendents and district assessment coordinators: The Colorado English Language Assessment (CELA) Placement Tests.* Retrieved September 17, 2006, from http://www.cde.state.co.us/cde_english/cela.htm

Medina, B. (2006b). *Memorandum: Colorado English language assessment (CELA) Placement Tests.* Retrieved July 11, 2006, from http://www.cde.state.co.us/cde_english/cela.htm

Test Administration Manual: See CTB/McGraw Hill (2007). Retrieved DATE, from http://www.ctb.com/

Colorado Dept. of Education website: http://www.cde.state.co.us/cde_english/cela.htm

Comprehensive English Language Learning Assessment (CELLA):

CELLA recommended cut scores. Retrieved September 14, 2006, from http://tennessee.gov/education/fedprog/doc/CELLACutScores806.doc *Comprehensive English language learning assessment*. Retrieved September 7, 2006, from http://tennessee.gov/education/fedprog/doc/fpcella.pdf

Florida Department of Education (January 2007). *Comprehensive English language learning assessment (CELLA): Florida CELLA regional train-the-trainer training session.* Retrieved March 11, 2007, from http://www.firn.edu/doe/aala/pdf/tttpresentation.pdf

Florida Department of Education (2006a). *Florida CELLA fact sheet.* Retrieved March 11, 2007, from http://www.firn.edu/doe/aala/pdf/cellainfosheet.pdf

Florida Department of Education (2006b). *Florida CELLA list of allowable accommodations.* Retrieved November 19, 2006, from http://www.firn.edu/doe/aala/pdf/allow_accom.pdf

Florida Department of Education (2005a). *Florida comprehensive English language learning assessment (CELLA)*. Retrieved November 19, 2006, from http://www.firn.edu/doe/aala/cella.htm

RETURN to TABLE of CONTENTS

Florida Department of Education (2005b). *Final response to USDE monitoring Title III.* Retrieved September 14, 2006, from http://www.fldoe.org/NCLB/pdfs/FRTIIIMR.pdf

Saavedra, L. (2005). *Comprehensive English language learning assessment*. Retrieved September 14, 2006, from http://www.firn.edu/doe/omsle/pdf/cella.pdf

Tennessee Department of Education. *ELL students and Tennessee assessments*. Retrieved September 14, 2006, from http://tennessee.gov/education/fedprog/doc/fp_TESTPolicy_Spring_06.pdf

Dakota English Language Proficiency Assessment (Dakota ELP):

Cook, G. (2005). *Aligning English language proficiency tests to English language learning standards.* Washington, D. C.: Council of State School Officers.

Ortman, D. (2005). *Dakota ELP assessment*. Retrieved September 18, 2006, from http://doe.sd.gov/oess/title/IIIela/docs/SDELP.pdf

Overview of the South Dakota assessment system. Retrieved September 21, 2006, from http://doe.sd.gov/octa/assessment/docs/SDassessment.pdf

English Language Development Assessment (ELDA):

A.I.R. (2005). English language proficiency standards and test and item specifications for grades 3–12. Report submitted to Council of Chief State School Officers (CCSSO) on behalf the LEP-SCASS by the American Institutes for Research (AIR) on October 31, 2005. Retrieved from: http://www.ccsso.org/projects/ELDA/Research_Studies/

Kopriva, R., Wiley, D. E., Chen, C-S., Levey, R., Winter, P. C., & Corliss, T. (2004). *Field test validity study results: English language development assessment final report.* Center for the Study of Assessment Validity and Evaluation (C-SAVE), University of Maryland College Park.

Malagon, M. H., Rosenberg, M. B., & Winter, P. C. (2005). *Developing aligned performance level descriptors for the English Language Development Assessment K–2 Inventories.* Council of Chief State School Officers (CCSSO). Retrieved September 29, 2006, from http://www.ccso.org/publications

Ferrara, S. and Sewell, D. (2006). *Design, psychometric, and instructional considerations for the speaking proficiency component of the English Development Language Assessment (ELDA).* Council of Chief State School Offices (CCSO), April 8, 2006. Retrieved September 29, 2006, from http://www.ccso.org/publications

SCASS-CCSSO (2005*). English Language Development Assessment district coordinator manual.* State Collaborative on Assessment and Student Standards (SCASS) in conjunction with American Institutes for Research. Retrieved November 11, 2006, from http://wvconnections.k12.wv.us/assessment.html

English Language Development Assessment (ELDA) grades 3–12:

AIR (2005a) English Language Development Assessment (ELDA) technical report: 2004 field test administration. Report submitted to Council of Chief State School Officers on behalf of the LEP-SCASS by the American Institutes for Research on October 31, 2005.

AIR (2005b*). English language proficiency standards and test and item specifications for grades 3–12.* Report submitted to Council of Chief State School Officers (CCSSO) on behalf the LEP-SCASS by the American Institutes for Research (AIR) on October 31, 2005. Retrieved from http://www.ccsso.org/projects/ELDA/Research_Studies/

Bunch, M. (2006) *Final report on ELDA standard setting*. Report submitted to the Council of Chief State School Officers by Measurement Incorporated on February, 2006. Retrieved from, http://www.ccsso.org/content/pdfs/ELDAStandardSettingFinalReport2005.pdf

Kopriva, R., Wiley, D. E., Chen, C-S., Levey, R., Winter, P. C., & Corliss, T. (2004*). Field test validity study results: English language development assessment final report.* Center for the Study of Assessment Validity and Evaluation (C-SAVE), University of Maryland College Park.

Malagon, M. H.; Rosenberg, M. B., & Winter, P. C. (2005). *Developing aligned performance level descriptors for the English Language Development Assessment K–2 Inventories.* Council of Chief State School Officers (CCSSO). Retrieved September 29, 2006, from http://www.ccso.org/publications .

RETURN to TABLE of CONTENTS

Ferrara, S. & Sewell, D. (2006). *Design, psychometric, and instructional considerations for the speaking proficiency component of the English Development Language Assessment (ELDA)*. Council of Chief State School Offices (CCSO), April 8, 2006. Retrieved September 29, 2006, from http://www.ccso.org/publications

SCASS-CCSSO (2005). *English Language Development Assessment district coordinator manual*. State Collaborative on Assessment and Student Standards (SCASS) in conjunction with American Institutes for Research. Retrieved November 11, 2006, from http://wvconnections.k12.wv.us/assessment.html

Idaho English Language Assessment (IELA):

Idaho State Board of Education (2006). *IELA Score Reports and Interpretation Guide*. Retrieved September 22, 2006, from http://www.boardofed.idaho.gov/lep/documents/06_InterpretiveGuide-v09.pdf

Idaho Department of Education (2006a). *2006 IELA Score reports interpretation guide*. Retrieved September 22, 2006, from http://www.boardofed.idaho.gov/lep/LEPAssessment.asp).

Idaho State Board of Education (2006b). *State Board of Education Meeting Minutes, Tab 2: Development standards, limited English proficiency program accountability plan ISAT, and IELA cut scores.* Retrieved September 22, 2006, from http://www.boardofed.idaho.gov/meetings/2006/SpecialMeetings/11_01_06/SBOE_Nov1-06.pdf

IPT® Title III Testing System (IPT):

Alaska Department of Education (2006). *Guide to test interpretation for the IDEA Proficiency Test (IPT) English Language Proficiency Assessment for parents and students, Spring 2006*. Retrieved October 29, 2006 from the Alaska Department of Education website: http://www.eed.state.ak.us/tls/assessment/elp.html

Bailey, A. L., & Butler, F. A. (2002). *An Evidentiary Framework for Operationalizing Academic Language for Broad Application to K-12 Education: A Design Document (CSE Tech. Rep. No. 611)*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ballard & Tighe (2004) *IPT Language Proficiency Progress Report*. Downloaded on October 15, 2006: http://www.nclb.ballard-tighe.com/

Del Vecchio, A. & Guerrero, M. (1995). Handbook of English language proficiency tests. Retrieved 8/22/06: http://www/mce;a/gwu.edu/pubs/eacwest/elptests.htm

North Carolina Department of Public Instruction (2005). Frequently asked questions (FAQs) – IDEA Proficiency Test IPT

IPT® 2004: IPT Early Literacy Test Reading and Writing (IPT Early Literacy Test, IPT Early Literacy R & W):

Massachusetts Department of Education (2007). *Principal's Administration Manual: IDEA Proficiency Test (IPT) Spring 2007*. Retrieved May 8, 2007, from http://www.doe.mass.edu/mcas/mepa/2007/IPTpam.pdf[8]

Kansas English Language Proficiency Test (KELPA):

Center for Educational Testing and Evaluation (2006). *Kansas English language proficiency assessment (KELPA) score report guide*. University of Kansas: Lawrence, Kansas. Retrieved January 25, 2007, from http://www.3.ksde.org/sfp/esol/kelpa_report_guide_final_spet_18_06.pdf

Kansas English language proficiency assessment KELPA. *Fall assessment conference*. Retrieved January 25, 2007 from http://www.ksde.org/LinkClick.aspx?fileticket=WyCgVrxDJ0M%3d&tabid=1636&mid=3450

Kansas State Department of Education (2007a). *Curricular standards for English to speakers of other languages*. Retrieved January 26, 2007, from http://www.ksde.org/Default.aspx?tabid=1636.

Kansas State Department of Education (2007b). *Kansas English language proficiency assessment (KELPA) questions and answers*. Retrieved January 26, 2007, from http://www.eslminiconf.net/katesol?Q&Akelpa.htm.

Kansas English language proficiency assessment KELPA (Spring 2006). *Score Report Guide*. Retrieved on March 30, 2007, from http://www.ksde.org/

[8] **Note:** this resource is in the process of being removed from the MA DOE Web site and will no longer be available at this URL.

RETURN to TABLE of CONTENTS

LinkClick.aspx?fileticket=RHzEwSjs4i4%3d&tabid=1636&mid=3450

KELPA testing window (2007). Retrieved March 30, 2007, from http://www.ksde.org/LinkClick.aspx?fileticket=hvq1eRzBV%2f0%3d&tabid=1636&mid=3450

Language Assessment Systems Links (LAS Links):

CTB McGraw-Hill. (2007). www.ctb.com/

Del Vecchio, A. & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque, NM. Evaluation Assistance Center: Western Region, New Mexico Highlands University. Retrieved from, www.ncela.gwu.edu/pubs/eacwest/elptests.htm

Jackson, S.L., Jackson, L.G. (2003, January) *Past, present, and future LAS results for program evaluation and student progress.* Paper presented at the National Association on Bilingual Education, New Orleans, LA.

Maculaitis Assessment of Competencies II (MAC II):

MO 2006 training AM session. Retrieved September 25, 2006, from http:// www.mo-mell.org/MACII.htm

TASA Literacy Online. *The MAC II test of English language proficiency*. Retrieved September 25, 2006, from http://www.tasaliteracy.com/mac/mac-main.html

Massachusetts English Language Assessment-Oral (MELA-O):

George Washington University (a). What is the Massachusetts English language assessment-oral or MELA-O? Retrieved February 5, 2007, from http://www.gwu.edu/~eaceast/projects/mela

George Washington University (b). *When was the MELA – 0 developed?* Retrieved February 5, 2007, from http://www.gwu.edu/~eaceast/projects/mela/history.html

George Washington University (c). *MELA-O pilot studies*. Retrieved February 5, 2007, from http://www.gwu.edu/~eaceast/projects/mela/pilot.html

Massachusetts Department of Education (2006a). Preliminary statewide results: Massachusetts English proficiency assessment (MEPA) spring 2006. Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/mepa/2006/results/prelim_s06state.doc

Massachusetts Department of Education (2006b). *Requirements for participation in of students with limited English language proficiency MCAS and MEPA*. Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/participation/lep.pdf

Massachusetts Department of Education (2005). *Massachusetts comprehensive assessment system: Important MEPA updates*. Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/2005/news/0819mepaupdate.html

Massachusetts Department of Education (2004a). *Massachusetts comprehensive assessment system: Statewide assessments for limited English proficient (LEP) students in 2004–2005*. Retrieved February 7, 2007, from http://www.doe.mass.edu/mcas/2004/news/0820epa.html

Massachusetts Department of Education (2004b). *Massachusetts Comprehensive Assessment System: Massachusetts English proficiency assessment (MEPA) standard-setting panels.* Retrieved February 5, 2007, from http://www.doe.mass.edu/mcas/2004/news/1206mepa.html

Viator, K., Dwyer, P. K., Wiener, D., & Meyer, A. (2006). *Spring 2006 MEPA administration workshop.* Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/mepa/06admin_wkshp.pdf

Massachusetts English Proficiency Assessment–Reading & Writing (MEPA-R/W):

Massachusetts Department of Education (2005). *Massachusetts comprehensive assessment system: Important MEPA updates*. Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/2005/news/0819mepaupdate.html

Massachusetts Department of Education (2006a). *Preliminary statewide results: Massachusetts English proficiency assessment (MEPA) spring 2006*. Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/mepa/2006/results/prelim_s06state.doc

Massachusetts Department of Education (2006b). *Requirements for participation in of students with limited English language proficiency MCAS and MEPA*. Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/2006/news/lep_partreq.pdf

Massachusetts Department of Education (June 2006). *Guide to Interpreting the 2006 MEPA reports for schools and districts*. Retrieved March 29, 2007, from http://

RETURN to TABLE of CONTENTS

www.doe.mass.edu/mcas/mepa/2006/interp_guide.pdf

Viator, K., Dwyer, P. K., Wiener, D., Meyer, A. (2006). *Spring 2006 MEPA administration workshop.* Retrieved September 25, 2006, from http://www.doe.mass.edu/mcas/mepa/06admin_wkshp.pdf

Montana Comprehensive Assessment System English Proficiency Assessment (MontCAS ELP):

Instruction, O. o. P., & Website, M. s. O. S. (2006a). *MontCAS English language proficiency assessment test coordinator's guide.* Retrieved January 26, 2007, from http://www.opi.mt.gov/pdf

Instruction, O. o. P., & Website, M. s. O. S. (2006b). *MontCAS English language proficiency assessment: Training for the fall 2006 Administration.* Retrieved January 26, 2007, from http://www.opi.mt.gov/PDF/Assessment/ELP/06ELPTraining.pdf

New Mexico English Language Proficiency Assessment (NMELPA):

Case, B. J. (2006). *New Mexico English Language Proficiency Assessment (NMELPA) accommodations for students with disabilities*. Harcourt Assessment, San Antonio Texas: April, 2006. Retrieved November 6, 2006, from the New Mexico Public Education Department web site: www.state.nm.us

New Mexico State Department of Education (2003a) *Frequently asked questions about English language learner student assessment and accommodations*. State of New Mexico. Retrieved November 6, 2006, from the New Mexico Public Education Department web site: www.state.nm.us

New Mexico State Department of Education (2003b) *New Mexico English language proficiency assessment (NMELPA) test coordinator's manual*. State of New Mexico. Retrieved November 6, 2006, from the New Mexico Public Education Department web site: www.state.nm.us

Consolidated state report (March 2006). Retrieved March 31, 2007, from http://www.ped.state.nm.us/div/learn.serv/Bilingual/l/title_III_reports/Consolidated%20State %20Report%20March%20200.doc

Ohio Test of English Language Acquisition (OTELA):

Ohio Department of Education (2006). *Guide to understanding test score results.* Retrieved December 10, 2006, from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?Page=3&TopicRelationID=1086&Content=22435

Texas Education Agency (2006a). *Texas English language proficiency assessment system (TELPAS) scoring manual.* Retrieved from http://www.tea.state.tx.us/student.assessment/resources/guides/interpretive/TELPA S_06.pdf#xml=http://www.tea.state.tx.us/cgi/texis/webinator/search/xml.txt?query=Texas+English+Language+Proficiency+Assessment+System&db=db&id=40a9dad8b839a213

Texas Education Agency (2006b). *Texas English language proficiency assessment system (TELPAS)*. Retrieved from http://www.tea.state.tx.us/student.assessment/resources/guides/coormanual/telpas06.pdf#xml=http://www.tea.state.tx.us/cgi/texis/webinator/search/xml.txt?query=Texa s+English+Language+Proficiency+Assessment+System&db=db&id=a890725020c9a8c0

Texas Education Agency (2006c). Texas English language proficiency assessment system (TELPAS). *Reading proficiency test in English: Test administration manual grades 2–12.* Retrieved from http://www.tea.state.tx.us/student.assessment/resources/guides/test_admin/2006/rpte_TAM.pdf

Texas Education Agency (2006d). Texas English language proficiency assessment system (TELPAS). *Texas observation protocols: Rater manual grades K–12.* Retrieved from http://www.tea.state.tx.us/student.assessment/admin/rpte/rater_manual_06.pdf

Texas Education Agency (2006e). *Chapter 4: Texas English language proficiency assessment system (TELPAS).* Retrieved from http://www.tea.state.tx.us/student.assessment/resources/techdig05/chapter4.pdf

Texas English Language Proficiency Assessment System (TELPAS) Other References

Texas Education Agency (2003). *Student Assessment Division: Technical Digest 2002–2003.* Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig/index.html

RETURN to TABLE of CONTENTS

Texas Education Agency (2004). *Student Assessment Division: Technical Digest 2003–2004*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig04/index.html

Texas Education Agency (2005). *Student Assessment Division: Technical Digest 2004–2005*. Retrieved March 30, 2007, from http://www.tea.state.tx.us/student.assessment/resources/techdig05/index.html

Texas Assessment (2006a). *Student assessment division: Technical digest 2005–2006*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Assessment (2006b). Student Assessment Division: Technical Digest 2005–2006. *Appendix 6*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/tx_dist_publ.htm

Texas Assessment (2006c). Student Assessment Division: Technical Digest 2005–2006. *Chapter 15: Validity*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/Chapters/Chapter15_Validity.pdf

Texas Assessment (2006d). Student Assessment Division: Technical Digest 2005–2006. *Chapter 17: Reliability*. Retrieved March 30, 2007, from http://k12testing.tx.ncspearson.com/TechDigest/Chapters/Chapter14_Reliability.pdf

Texas Education Agency (2007). Spring 2007 TOP Rater Manual. Retrieved August 26, 2007 from: http://www.tea.state.tx.us/student.assessment/admin/rpte/TP07_TOP_Rater_Man ual_Final_tagged.pdf

Oregon English Language Proficiency Assessment (ELPA):

Accommodations Table. Retrieved September 24, 2006 from http://www.ode.state.or.us/search/page/?id=487

ELPA Frequently Asked Questions September 1, 2005. Retrieved September 24, 2006 from http://www.ode.state.or.us/teachlearn/testing/admin/ell/asmtelpafaq050901.pdf

ELPA Overview September 1, 2005. Retrieved September 24, 2006 from http://www.ode.state.or.us/teachlearn/testing/admin/ell/asmtelpaoverview050901.pdf

ELPA Item Types–Instruction Transcripts. *Item type: Reading/multiple choice*. Retrieved September 24, 2006 from http://www.oregonelp.net/oregonelp/resources/ode_images/ELPA_item_guide.pdf

ELPA test administration proctor guide. Retrieved September 24, 2006 from http://www.oregonelp.net/oregonelp/resources/ode_images/ELPA_Proctor_Guide.pdf

Language Learners Solution (2005). Oregon Department of Education awards ELP assessment contract to language learning solutions. Retrieved September 24, 2006 from http://www.onlinells.com/news.php#oregon

Letter written to Superintendents, Principals, District Test Coordinators , Others in regards to Oregon Assessment System–New Achievement Standards. *How will Oregon set new achievement standards?* Retrieved on March 29, 2007 from http://www.ddouglas.k12.or.us/downloads/documents/DD_Setting new Performa

Oregon Department of Education (2006). *ELPA training guide*. Retrieved September 24, 2006 from http://www.oregonelp.net/oregonelp/resources/ode_images/ELPA_Training_Guide_051706.doc

Oregon English Language Proficiency Assessment (ELPA). *K–1, 2–3 Frequently Asked Questions*. Retrieved September 24, 2006 from http://www.ode.state.or.us/teachlearn/testing/admin/ell/asmtelpak1-23faq.pdf

Oregon Department of Education (2007): *Oregon English Language Proficiency Assessment Users Manual*. Retrieved July 2, 2007 from: http://www.ode.state.or.us/teachlearn/testing/admin/ell/elpausersmanual.pdf

Oregon Department of Education: Prepared by Carter, S. *K–2 English language proficiency assessment (ELPA) issues and recommendations executive summary*. Retrieved September 24, 2006 from http://www.ode.state.or.us/teachlearn/testing/admin/ell/asmtelpak-2execsummresrecom.pdf

Oregon Department of Education. *Assessment Update 3, (2) August 28, 2006*. Retrieved September 24, 2006 from http://www.ode.state.or.us/teachlearn/testing/asmtupdate08282006.pdf

Super (2005–2006) AMAO Report (Public Release on Friday, 11/3/06) Retrieved March 29, 2007 from http://listsmart.osl.state.or.us/pipermail/super/2006–November/000503.html

RETURN to TABLE of CONTENTS

The Stanford English Language Proficiency Test (SELP, Stanford ELP):

Barnett, J. (2005). *The Stanford English language proficiency test.* Retrieved from http://www.mde.k12.ms.us/acad/OSA/ELP_agenda.doc.

Cook, G. (2006). *Findings of an alignment study of the Stanford English language proficiency* test and the English language proficiency (World-class instructional design and *assessment) to the Delaware grade level expectations in English language arts for grades 3, 4, 5, 6, 7, 8, 9, and 10.* Submitted to the Assessment and Accountability Branch of the Delaware Department of Education. Retrieved July 5, 2007, from http://www.doe.k12.de.us/aab/DE%20SELP-ELP%20Alignment%20Report--Final.pdf

Harcourt Assessment. (2007). *Stanford English language proficiency test and Stanford Spanish language proficiency test technical manual (2nd ed.).* San Antonio, TX.

Mississippi Department of Education (2005). *ELL updates.* Retrieved from http://www.mde.k12.ms.us/ACAD/osa/04_05_reminders_for_ell.pdf#search=%22Stanford%20English%20Language%20Proficiency%20Tests%22

Mississippi Department of Education (2003). *Frequently asked questions about English language learners (ELLs).* Retrieved from http://www.mde.k12.ms.us/acad/osa/ELL_FAQ.pdf#search=%22Stanford%20English%20Language%20Proficiency%20Tests%22

Question and Answer Document. Retrieved September 25, 2006, from http://www.doe.virginia.gov/VDOE/Instruction/ESL/SELP_FAQ.pdf

The Stanford English language proficiency test. Harcourt Assessment. Retrieved November 25, 2006, from http://harcourtassessment.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8429-206&Mode=summary

Test of Emerging Academic English (TEAE):

Albus, D., Klein, J. A. Liu, K., & Thurlow, M. (2004). *Connecting English language proficiency, statewide assessments and classroom proficiency* (LEP Projects Report 5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 20, 2006, from http://education.umn.edu/NCEO/OnlinePubs/LEP5.html

Minnesota Department of Education (2003). *An alignment study of the test of emerging academic English and Minnesota's grade level expectations for reading, grades 3, 5, 7 and 10.* St. Paul, MN: Division of Statewide Assessment and Testing,

Minnesota Department of Education. Retrieved September 9, 2004, from http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/

Minnesota Department of Education (2003). Test of Emerging Academic English (TEAE) standards setting report: summer 2002–winter 2003. Retrieved September 9. 2006, from: http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/ELL_Tests/ELL_Technical_Reports/index.html

Minnesota Department of Education (2004). *An alignment study of the test of emerging academic English and Minnesota's English language proficiency standards.*

St. Paul, MN: Division of Statewide Assessment and Testing, Minnesota Department of Education. Retrieved September 9, 2005, from http://education.state.mn.us/mde/Accountability_Programs/Assessment_and_Testing/

Texas English Language Proficiency Assessment System (TELPAS):

Interpretive Guide (2006). Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/resources/guides/interpretive/TELPAS_06.pdf#xml=http://www.tea.state.tx.us/cgi/texis/webinator/search/xml.txt?query=Texas+English+Language+Proficiency+Assessment+System&db=db&id=40a9dad8b839a213

Texas Student Assessment Coordinator Manual (2006). Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/resources/guides/coormanual/telpas06.pdf#xml=http://www.tea.state.tx.us/cgi/texis/webinator/search/xml.txt?query=Texas+English+Language+Proficiency+Assessment+System&db=db&id=a890725020c9a8c0

TOP Proficiency Level Descriptors. (2006). Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/admin/rpte/TOP_pld.pdf

RETURN to TABLE of CONTENTS

TOP Rater Manual Grades K–12. (2006) Retrieved September 26, 2006, from http://www.tea.state.tx.us/student.assessment/admin/rpte/rater_manual_06.pdf

Utah Academic Language Proficiency Assessment (UALPA):

Utah State Office of Education (2006). *UALPA coordinator's manual*. Retrieved March 29, 2007, from http://www.usoe.k12.ut.us/Eval/DOCUMENTS/UALPA_Coordinators_Manual.pdf

Utah State Office of Education (2005). *Curriculum & Instruction Assessment & Accountability Alternative Language Services.* PowerPoint presented at the Directors Meeting on November 17, 2005. Retrieved March 29, 2007, from http://www.usoe.k12.ut.us/Eval/DOCUMENTS/ELL_Meeting.ppt

Utah State Office of Education (October 2006). *The Utah academic language proficiency assessment UALPA.* Retrieved March 29, 2007, from http://www.usoe.k12.ut.us/Eval/DOCUMENTS/UALPA_Training.ppt#258,1,The

Virginia Stanford English Language Proficiency Test (VASELP):

Virginia Stanford English Language Proficiency (SELP) Test (2007). *Test implantation manual.* Retrieved March 31, 2007, from http://www.pen.k12.va.us/VDOE/Assessment/SELP/VA07_SELP_TIM.pdf

Detailed Speaking Rubric (N.D.). Retrieved September 25, 2006, from http://www.k12.wy.us/FP/title3/Detailed_Speaking_Rubric.pdf

Harcourt (2005). *Wyoming SELP overview*. Retrieved September 25, 2006, from http://www.k12.wy.us/FP/title3/SELP_Overview.pdf

Harcourt (2004a). *Scoring the SELP speaking test*. Retrieved September 25, 2006, from http://www.k12.wy.us/FP/title3/selp_speak_scoring.pdf

Harcourt (2004b). *Scoring the SELP writing test.* Retrieved September 25, 2006, from http://www.k12.wy.us/FP/title3/selp_write_scoring.pdf

Washington Language Proficiency Test II (WLPT-II):

Harcourt (2006). *WLPT-II Handouts*. Retrieved September 25, 2006, from http://www.k12.wa.us/assessment/pubdocs/WLPTHandouts.pdf

Harcourt (2006). *WLPT II Overview*. Retrieved September 25, 2006, from http://www.k12.wa.us/assessment/pubdocs/WLPTOverview.ppt

Harcourt (2004). *Scoring the speaking subtest*. Retrieved September 25, 2006 from http://www.k12.wa.us/assessment/pubdocs/ScoringWLPTSpeaking.ppt

Washington Language Proficiency Test II. *Test Coordinator Manual*. Retrieved September 25, 2006, from http://www.k12.wa.us/Assessment/TestAdministration/pubdocs/WLPTTCM_2006.pdf

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007

State	First Implemented	Name of Test	Test Developer
Alabama	Spring 2005	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Alaska	Spring 2006	IPT® Title III Testing System (IPT) [LAS (Forte, 2007)]	Ballard & Tighe
Arizona	Fall 2006	Arizona English Language Learner Assessment (AZELLA)	Arizona Department of Education; Harcourt Assessment Inc.

Arkansas	Spring 2007	English Language Development Assessment (ELDA)
		English Language Development (ELDA) K–2 Assessment
			American Institute for Research (AIR);
			Center for the Study of Assessment Validity and Evaluation (C-SAVE);
			Council of Chief State School Officers (CCSSO); Measurement Inc.(MI);
			State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)

California	Fall 2001	California English Language Development Test (CELDT)	California Department of Education; CTB/McGraw Hill
Colorado	Spring 2006	Colorado English Language Assessment (CELA)	CTB/McGraw Hill
Connecticut	Winter & Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Delaware	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Florida	Fall 2006	Comprehensive English Language Learning Assessment (CELLA)	Accountability Works; Educational Testing Service (ETS); and a consortium of 5 states
Georgia	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Hawaii	Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Idaho	Spring 2006	Idaho English Language Assessment (IELA)	Questar Assessment, Inc. (formerly Touchstone Applied Science Associates)
Illinois	Spring 2006	Assessing Comprehension and State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Indiana	Winter and Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill

Iowa	Spring 2006	English Language Development Assessment (ELDA)
		English Language Development (ELDA) K–2 Assessment
			American Institute for Research (AIR);
			Center for the Study of Assessment Validity and Evaluation (C-SAVE);
			Council of Chief State School Officers (CCSSO); Measurement Inc.(MI);
			State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
Kansas	Spring 2006	Kansas English Language Proficiency Assessment (KELPA)	The Center for Testing and Evaluation (CETE); Kansas State Department of Education; University of Kansas
Kentucky	Spring 2007	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Louisiana	1. Spring 2005 (grades 3-12) 2. Spring 2006 (grades K-2 added)	1. English Language Development Assessment (ELDA) 2. English Language Development (ELDA) K–2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Maine	Spring 2005	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Maryland	Spring 2006	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
Massachusetts	1. Spring 2004 2. Spring 2004 3. Spring 2007	1. Massachusetts English Proficiency Assessment-Reading & Writing (MEPA-R/W) 2. Massachusetts English Language Assessment-Oral (MELA-O) 3. IPT® 2004:IPT Early Literacy Test reading and writing (K-2 reading and writing only)	1. Massachusetts Department of Education; Measured Progress 2. Educational Assistance Center (EAC) East; Massachusetts Assessment Advisory Group (MAAG); Massachusetts Department of Education 3. Ballard & Tighe
Michigan	2006	Michigan English Language Proficiency Assessment (MI-ELPA)	Harcourt Assessment Inc.; Michigan Department of Education
Minnesota	1. Fall 2001 2. 2002 – 2003 academic year	1. Test of Emerging Academic English (TEAE) 2. MN SOLOM	1.MetriTech, Inc.; Minnesota Department of Education 2. Bilingual Education Office of the California Department of Education; San Jose Area Bilingual Consortium
Mississippi	2003 - 2004 academic year	The Stanford English Language Proficiency Test (SELP, Stanford ELP)	Harcourt Assessment Inc.
Missouri	Winter 2002	Maculaitis Assessment of Competencies II (MAC II)	Questar Assessment, Inc. (formerly Touchstone Applied Science Associates)
Montana	Winter 2006	MontCAS English Language Proficiency Assessment (MONTCAS ELP)	Measured Progress; Mountain West Assessment Consortium (MWAC); Questar Assessment, Inc. (formerly Touchstone Applied Science Associates) has taken over production of test

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
Nebraska	Spring 2005 (grades 3-12) Spring 2006 (grades K-2 added)	English Language Development Assessment (ELDA) English Language Development (ELDA) K–2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Nevada	2005 - 2006 academic year	Language Assessment System Links (LAS Links)	CTB/McGraw Hill
New Hampshire	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
New Jersey	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
New Mexico	Spring 2006	New Mexico English Language Proficiency Assessment (NMELPA)	Harcourt Assessment Inc.; New Mexico Department of Education
New York	Spring 2005	New York State English as a Second Language Achievement Test (NYSESLAT)	Educational Testing Service (ETS); Harcourt Assessment Inc.; New York State Education Department
North Carolina	2005	IPT® Title III Testing System (IPT)	Ballard & Tighe
North Dakota	Spring 2007	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Ohio	1. Spring 2006 (grades 3-12) 2. Spring 2006 (K-2 only)	1. Ohio Test of Language Acquisition (OTELA) 2. English Language Development (ELDA) K-2 Assessment	1. American Institutes for Research (AIR); Ohio Department of Education 2. American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Oklahoma	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Oregon	Spring 2006	Oregon English Language Proficiency Assessment (ELPA) [SELP (Forte, 2007)]	Language Learning Solutions (LLS)
Pennsylvania	Spring 2007	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
Rhode Island	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
South Carolina	Spring 2005 (Grades 3-12) Spring 2006 (Grades K-2 added)	English Language Development Assessment (ELDA) English Language Development (ELDA) K-2 Assessment	American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
South Dakota	Spring 2006	Dakota English Language Proficiency Assessment (Dakota ELP)	Harcourt Assessment Inc.; South Dakota Department of Education
Tennessee	1. Spring 2005 2. 2007 3. 2007	1. Comprehensive English Language Learning Assessment (CELLA) 2. English Language Development Assessment (ELDA) 3. English Language Development (ELDA) K-2 Assessment	1. Accountability Works; Educational Testing Service (ETS); and a consortium of 5 states 2./3. American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Texas	 A. 2000 B. 2005	1. Texas English Language Proficiency Assessment System (TELPAS) A. Reading Proficiency Test in English (RPTE) B. Texas Observation Protocols (TOP)	Beck Evaluation and Testing Associates (BETA); Pearson Educational Measurement; Texas Education Agency (TEA)
Utah	Fall 2006	Utah Academic Language Proficiency Assessment (UALPA)	Measured Progress; Mountain West Assessment Consortium (MWAC)
Vermont	Spring 2005	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Virginia	Spring 2006	Virginia Stanford English Language Proficiency Test	Harcourt Assessment, Inc.
Washington	2006	Washington Language Proficiency Test II (WLPT-II)	Harcourt Assessment Inc.; Washington Department of Education
Washington D.C.	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

TABLE 1. Tests Currently Used by States for Title III Reporting Purposes by State, as of August 2007 (cont.)

State	First Implemented	Name of Test	Test Developer
West Virginia	1. 2005 2. Spring 2005 (grades 3-12) 3. Spring 2006 (grades K-2 added)	1. West Virginia Test for English Language Learning (WESTELL) 2. English Language Development Assessment (ELDA) 3. English Language Development (ELDA) K-2 Assessment [ELDA only (Forte, 2007)]	1. N/A 2./3. American Institute for Research (AIR); Center for the Study of Assessment Validity and Evaluation (C-SAVE); Council of Chief State School Officers (CCSSO); Measurement Inc.(MI); State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS)
Wisconsin	Spring 2006	Assessing Comprehension and Communication State to State for English Language Learners (ACCESS for ELLs®, ACCESS)	Center for Applied Linguistics (CAL); World-Class Instructional Design and Assessment Consortium (WIDA)
Wyoming	Spring 2006	Wyoming English Language Learner Assessment (WELLA)	Harcourt Assessment Inc.; Wyoming Department of Education

Note: *Data on the states' current ELP assessments that were obtained by this study were compared with similar data provided in Forte (2007). The data from the two sources are generally consistent. In a few cases, minor discrepancies are provided from both sources.*

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

RETURN to TABLE of CONTENTS

Strengthening Teaching and Learning for All

UCDAVIS
School of Education

School of Education
UC Davis
One Shields Ave
Davis, CA 95616
http://education.ucdavis.edu/