World-Class Instructional Design and Assessment

Alternate Access for ELLs™
Series 100 Development and Operational Field Test: Technical Report

Draft

Prepared by:

CAL/WIDA Partnership Activities
Psychometrics/Research Team

Center for Applied Linguistics

06/03/2013

Contents

1 Description of Alternate ACCESS for ELLs™

1.1 Purpose of this report

The purpose of this Technical Report is to provide technical information with regard to the development and technical analysis of the field test for the Alternate ACCESS for ELLs™ test forms (Series 100), hereafter referred to as Alternate ACCESS. The technical information herein is intended for use by those who have technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999).

In this report, empirical data from the field test administration for Alternate ACCESS were used to provide initial technical information about the assessment. Data analyses using the population data from the operational administration for Alternate ACCESS had been planned in order to provide complete technical information about the assessment.

1.2 Overview of Assessment

Alternate ACCESS is an assessment of English language proficiency (ELP) for students in Grades 1–12 who are classified as English language learners (ELLs) and who have significant cognitive disabilities that prevent their meaningful participation in the ACCESS for ELLs® (Center for Applied Linguistics [CAL], 2012b), hereafter referred to as ACCESS. Alternate ACCESS is a first attempt made by World-Class Instructional Design and Assessment (WIDA) to assess ELP for ELLs with significant cognitive disabilities. As such, the assessment continues to be refined to clarify the construct and to develop test design which better reflects the variety of student language use in this population.

1.3 Organization of the WIDA English Language Development (ELD) Standards

1.3.1 Description of the WIDA ELD Standards

The design of Alternate ACCESS is built upon the foundational WIDA ELD standards. The five WIDA ELD standards are:

Standard 1—Social and Instructional Language (SIL)

ELLs communicate in English for **social and instructional** purposes in the school setting.

Standard 2— Language of Language Arts (LoLA)

ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Language Arts**.

Standard 3—Language of Mathematics (LoMA)

ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Math**.

Standard 4—Language of Science (LoSC)

ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Science.**

Standard 5—Language of Social Studies (LoSS)

ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Social Studies.**

1.3.2 Grade Level Clusters

Alternate ACCESS is organized into the same grade-level clusters as ACCESS: grades 1-2, 3-5, 6-8, and 9-12. A Kindergarten version of the test is not currently available, but is planned for future development.

1.3.3 The Four Domains

The WIDA ELD Standards describe developing ELP for each of four language domains: listening, reading, speaking, and writing. The Alternate ACCESS test includes a section to assess each of these four domains.

1.3.4 The Proficiency Levels

The WIDA Alternate ELD Standards describe growth in ELP over six levels. These six levels include three newly developed language proficiency levels and three levels that relate to the WIDA ELD Standards for the general population. The most basic proficiency level is Proficiency Level (PL) A1: 'Initiating,' and the most advanced stage of language proficiency described is PL 3: 'Developing'. The first three levels of the Alternate ELD PLs, A1 – A3, are language proficiency antecedents to the existing WIDA ELD PL 1for the general population. An important aspect of the Alternate ELD levels (A1-A3) is that they represent small chunks of language growth within the PL 1. A highlight of this structure is that progress in language acquisition for students with significant cognitive disabilities can be identified in smaller and narrower gradations. Figure 1A below presents a conceptualization of PL1, which has been stretched to include levels A1 – A3. PL2 and PL3 have not been stretched in such a way in this figure; however, these levels are as broad as PL1 (i.e., the difference between a PL1 and a PL2 item, as well as the difference between a PL2 and PL3 item, is roughly the same as the difference between an A1 and a PL1 item **in terms of difficulty**). The proficiency levels assessed in Alternate ACCESS are illustrated in Figure 1A.

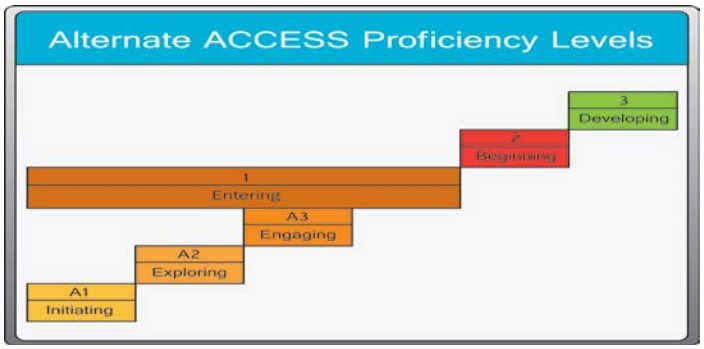

Figure 1A: Alternate ACCESS Proficiency Levels

1.3.5 Alternate Model Performance Indicators

The assessment of performance on the Alternate PLs (A1, A2, A3) is based on Alternate Model Performance Indicators (AMPIs), which follows the same paradigm as the Model Performance Indicators (MPIs) for each proficiency level defined by the WIDA ELD Standards in ACCESS. The AMPIs specify how ELLs with significant cognitive disabilities may use and process academic English language at a particular proficiency level within a grade level, language standard, and language domain.

Example: Grade 1-2, Social & Instructional Language, Listening: PL 1

> "Identify symbols, objects, or people associated with classrooms or school areas, personnel or activities from pictures and oral statements."

Test developers use the AMPIs to write tasks (i.e., test items) that allow students to demonstrate their developing proficiency in English.

1.4 Description of the Assessment

Alternate ACCESS is divided into grade clusters: 1-2, 3-5, 6-8, and 9-12. Each grade cluster test has individual sections for the four language domains of Listening, Reading, Speaking, and Writing. Alternate ACCESS was developed with the intention of allowing ELLs with significant cognitive disabilities to demonstrate their developing proficiency in English and as such has test features which reflect the different challenges of assessing this population. For example, students are given multiple opportunities to respond to test items in all sections of the test, which provides supports and ensures access to all students. Alternate ACCESS is a short test, with no individual section (i.e., Listening, Reading, Speaking, and Writing) containing more than 10 items. The number of test items was intentionally limited so that the test would not present undue stress to the students taking it.

Additional features of the test include simplified language; repetition of questions; heavy reliance on graphics rather than on text; larger size of testing materials, font, and graphics; and availability of cues and supplemental questions. During test administration, individualized instructional supports and accommodations that meet state-specific guidelines may be used. The

next section (1.5 Test Administration) of this document will go in to further detail about how test administration and individualized supports were incorporated into the assessment. The attributes of the assessment are summarized in Table 1A.

Table 1A
Alternate ACCESS for ELLs™

Grade-level clusters	1-2, 3-5, 6-8, and 9-12
Language domains	Listening, Reading, Speaking, and Writing
Task format	Selected Response (Listening and Reading) and Constructed Response (Speaking and Writing)
Standards	Social and Instructional Language (SIL); Language of Language Arts (LoLA); Language of Mathematics (LoMA); Language of Science (LoSC)
ELP Levels	A1:Initiating, A2: Exploring, A3: Emerging, P1: Entering, P2: Beginning, P3: Developing
Tasks based on	Alternate Model Performance Indicators for A1, A2, A3 Model Performance Indicators for P1, P2, P3
Administration	Individual
Scoring	Test administrator scores all sections

1.5 Test Administration

As mentioned in the preceding section, Alternate ACCESS is made up of four sections assessing the four language domains — Listening, Reading, Speaking, and Writing; the Speaking section is divided into two parts, Part A and Part B. The Writing section is divided into three parts, Parts A, B, and C. Both are described in further detail below. All sections of Alternate ACCESS are semi adaptive, which means that the administration of a test section may be ended if the student responds incorrectly or does not respond to three consecutive tasks. The Test Administrator Manual suggests that each section should take no longer than 20 minutes. However, different domain sections can be administered on different days with no minimum or maximum break between the administrations, as long as the entire Alternate ACCESS assessment is administered within the allotted testing window. In the Writing and Speaking sections, which contain multiple parts, the order of the parts cannot be changed (i.e., Part A needs to be administered before Part B). All sections of the test must be administered unless otherwise stipulated in the student's Individualized Education Program (IEP).

1.5.1 Listening and Reading Sections

The Listening and the Reading sections provide students with multiple opportunities to demonstrate their knowledge. The following steps are used to administer each task in the Listening and the Reading sections.

1. Administer CUE A (initial prompt and question of the task).
2. If the student does not respond, the test administrator must repeat CUE A again, as indicated in the test administrator's script.
3. If the student answers incorrectly or does not respond to CUE A, the test administrator will read CUE B. CUE B simplifies the initial prompt and asks the question again.
4. If the student responds incorrectly, or does not respond at all after the test administrator reads CUE B, the test administrator will administer CUE C. This cue provides the answer to the question, restates the prompt, and asks the question again.

Based on these administration guidelines for Listening and Reading, a student has a maximum of four opportunities to respond to each task (CUE A – 2, CUE B – 1, CUE C – 1). If a student responds correctly to the task at CUE A (including if the teacher repeated CUE A) the test administrator will score the task as **Correct at CUE A**. If after the two possible attempts at CUE A the test administrator moves on to CUE B and the student answers correctly, they will be scored as **Correct at CUE B**. Likewise, if the student has reached CUE C and answers correctly, they will be scored as **Correct at CUE C**. Finally, if after the four possible chances to answer the task the student has not selected the correct answer, the teacher will mark the task as **Incorrect**. If the student did not respond to any of the four opportunities, the task will be marked as '**No Response**.' Chapter 5.2 will describe how values were assigned to each possible score for a task in more detail.

1.5.2 Speaking Section

The Speaking section has two thematic folders, Parts A and B. Thematic folders are a set of tasks based on a common setting or story (e.g., students in the library). The graphic(s) and character(s) often remain the same for all the tasks in a thematic folder.

- Part A of the Speaking section has tasks at levels A1 - A3.
- Part B of the Speaking section has tasks at levels A1 - P2.
- The script for all tasks includes three questions (Question 1, 2, and 3) as multiple opportunities for the student to provide a response at a given task level.

In the Speaking section the student is given up to six opportunities to respond. This provides students with multiple opportunities to respond appropriately to the task in English. For each task, the test administrator reads Question 1 and prompts the student to respond. If the student does not score '**Meets**,' the test administrator must repeat the task again. If the student still does not score **'Meets'** after the repetition, the test administrator must ask Question 2, which simplifies the prompt and, in some tasks, models the expected response. If the student still does

not score '**Meets**,' Question 2 must be repeated, and if the student does not score **'Meets'** after that, the test administrator must administer Question 3. The additional questions can also be repeated to the students once. The possibility for repetition for all the Questions provides the student with six opportunities to produce a response in each Speaking task. If the student produces an appropriate response to the task at any point within six provided opportunities, the task is scored as '**Meets**.' If the student is not able to produce a response that meets task expectations, a score of **'Approaches'** is assigned. If the student does not make any attempt to respond to the task, a score of '**No Response'** is assigned. Chapter 5.2 will describe the value assigned to each possible score for a task in more detail. The Test Administration Manual (WIDA, 2012a) instructed teachers to score the Speaking section using scoring guidance provided in a column of the Speaking score sheet termed the 'Expect' box. For each task, the 'Expect' box provides the test administrator with a description of a response that would meet the task expectations (e.g., repeat a word or produce a phrase related to the task). The scoring guidelines in the 'Expect' boxes parallel the Speaking rubric available in the Test Administrator Manual. The Speaking rubric was developed by a team of CAL test developers and was based on the Performance Definitions for the Alternate ACCESS proficiency levels.

1.5.3 Writing Section

The Writing section has three thematic folders, Parts A, B, and C.

- Part A of the Writing section has tasks at levels A1- P1.
- Part B of the Writing section has tasks at levels A1 –P1.
- Part C provides the student with tasks at Levels P1 – P3; a student is only administered Part C if s/he scores '**Meets**' on 7 of the 8 tasks in Parts A and B.

In Parts A and B of the Writing section, the script is designed for the test administrator to model each task for the student. This provides students the opportunity to observe the test administrator perform the task before trying it. For example, in the first task of the Writing section, the test administrator's script will instruct the test administrator to draw a circle around an image before asking the student to do the same. Similar to the Speaking section, each task in the Writing section provides the student with multiple opportunities for the student to produce a response. If the student produces a response that is appropriate for the task, a score of **'Meets'** is assigned, and if the student does not produce a response that meets task expectations, a score of **'Approaches'** is assigned. If the student does not respond during the task administration, '**No Response'** is assigned to the task. The Test Administration Manual (WIDA, 2012a) instructed teachers to score the Writing section using scoring guidance provided in a column of the Writing score sheet termed the 'Expect' box. For each task in Parts A and B, the 'Expect' box provides the test administrator with a description of a response that would meet the task expectations (e.g., copy or write a word related to the task). The scoring guidelines in the 'Expect' boxes parallel the Writing rubric available in the Test Administrator Manual and the Student Response Booklet. Part C is scored based on the Writing rubric. Student performance can receive a score of 'Meets 1,' 'Meets 2,' 'Meets 3,' 'Approaches,' or 'No Response.' A score of 'Meets' 1, 2 or 3

corresponds to writing performances described in the Writing rubric for writing at PL 1, 2, or 3. The Writing rubric was developed by a team of CAL test developers and was based on the Performance Definitions for the Alternate ACCESS proficiency levels.

1.6 Criteria for participation

Eligibility for participation in the Alternate ACCESS test is determined by IEP teams. All the following criteria that must be met in order for a student to qualify for the test:

- The student is classified as an **ELL**.
- The student has a **significant cognitive disability** and receives special education services under the Individuals with Disabilities Education Act (2004).
- An IEP Team has determined that the student will participate in an **alternate curriculum** wherein:
    - Accommodations and modifications within the general education curriculum were considered.
    - The decision to participate in the alternate curriculum is not primarily due to social, cultural, or economic factors.
    - The student's curriculum more closely reflects the AMPIs than typical age- or grade-appropriate benchmarks.
- The student is or will be participating in his or her statewide **alternate accountability assessment**.

1.7 Scores

As with all WIDA assessments, scale scores on Alternate ACCESS are provided for each of the four language domains: Listening, Speaking, Reading, and Writing. In addition, composite scores are provided for Oral Language (Listening and Speaking), Comprehension (Reading and Listening), Literacy (Reading and Writing), and Overall (Reading, Writing, Listening, and Speaking). All Alternate ACCESS scale scores are in terms of the reporting scale that underlies Alternate ACCESS from 1 to 12, allowing users to compare scores from year to year as students progress through their educational experience.

2 Development of Alternate ACCESS for ELLs™

The initial grant phase development of Alternate ACCESS was initiated by a research group at the University of Wisconsin under the supervision of Dr. Craig Albers. The grant phase development work, partially funded by a U.S. Department of Education Enhanced Assessment Instruments Grant, occurred between 2008 and 2011.

The grant phase development team at the University of Wisconsin produced the AMPIs and preliminary field test forms for grades K-12. After the grant phase concluded, the project was handed off to WIDA in early 2011. In collaboration with WIDA, the Center for Applied Linguistics (CAL) began working toward the goal of producing operational test forms in the

spring of 2011. Staff members at WIDA and CAL worked together with experts on special education issues, including the assessment of severely cognitively disabled children, and met several times to identify, plan, and implement changes to the assessment in order to move the field test forms from their preliminary state to operational test forms.

The initial review of the field test forms by WIDA and CAL yielded two major outcomes: a) the Listening and Reading items would need significant revisions, and b) the Speaking and Writing sections for the test would need to be recreated. Based on the outcomes from this initial review, the following timeline details the steps that were taken to create the operational test forms:

- Speaking & Writing Workshop – August 1-3, 2011:
  To facilitate the redevelopment of the Speaking and Writing sections of the test, an item writing workshop was convened. The Speaking and Writing workshop brought ELL and special education experts to CAL to develop tasks. Over two days, educators used the Speaking and Writing AMPIs to guide their development of test tasks based around grade level appropriate themes.

- Test Revision – August - October, 2011:
  From August through October, 2011, CAL's test development team continued creating and refining new Speaking and Writing tasks from the workshop. Additionally, revisions identified during the initial review were made in all four test domains.

- Bias & Sensitivity Review – October 12-13, 2011
  The Bias and Sensitivity review included ELL and special education experts from across the WIDA Consortium. Bias and Sensitivity reviewers were given the task of determining whether the test items were in any way inappropriate in two ways: a) the item presented a potential source of bias for test takers, or b) the item was presenting information that could be disconcerting for a test taker based on any number of factors (e.g., age, race or ethnicity, religious background). During this review, experts discussed in detail each task in every section of the test for all grade clusters. Using the information from the two review activities, CAL's test development team made further revisions to all test sections. The list of participants in the Bias and Sensitivity review is in Table 2A below:

Table 2A
Bias and Sensitivity Review Participants

Name	State	Affiliation
Cheri Erdel	MO	Mexico, Missouri Public Schools
Connie Thiebeault	VA	Fairfax County Public Schools
Donna DeVito	IL	Cicero, Illinois Public Schools
Karen Timmons	DE	Frankford, Delaware Public Schools
Katy Decker	ND	Gwinner, North Dakota Public Schools
Lynne Angus-Barker	MS	Pascagoula, Mississippi
Rosemary Gardner	WI	Madison Public Schools
Sandy Berndt	WI	WIDA
Susan Beard	AL	Alabama State Department of Education

- Content Review – October 13-14, 2011:

  The Content review was held with ELL and special education experts from across the WIDA Consortium. Content reviewers were given the task of determining whether the test content was in any way inappropriate based on the accuracy of information presented and the grade level appropriateness of the task content. During this review, experts discussed each task in every section of the test for all grade clusters in detail. Using the information from the two review activities, CAL's test development team made further revisions to all test sections. The list of participants in the Content review is listed in Table 2B.

  Table 2B
  Content Review Participants

Name	State	Affiliation
Linda Rashidi	DC	Fairfax, Virginia Public Schools
Diane Bonney	ME	Regional School Unit 24, Maine Public Schools
Donna DeVito	IL	Cicero, Illinois Public Schools
Cheryl Gosnell	MO	California, Missouri Public Schools
Sandra Bryd	KY	Kentucky Public Schools
Iris Jacobson	WI	Oshkosh, Wisconsin Public Schools
Rosemary Gardner	WI	Madison, Wisconsin Public Schools
Sandy Berndt	WI	WIDA
Cheri Erdel	MO	Mexico, Missouri Public Schools
Deborah Howard	ME	Kennebuc, Maine Public Schools

- Pilot Testing – November 15-22 , 2011:

  Pilot testing for all grade-level clusters of Alternate ACCESS was conducted from November 15th – 22nd, 2011, and included eight schools in Maryland, Virginia, and Washington, DC. The pilot test had two major components: a) CAL researchers observed teachers administer the test form to their student, and b) CAL researchers interviewed

teachers after the administration to gather information about the process and clarity of test administration. Revisions to all test forms were made based on data gathered from the pilot testing, and the forms were then sent to a professional copy editor. A list of the schools and teachers that participated in the Pilot Test is shown in Table 2C.

Table 2C

Schools and Teachers that Participated in the Pilot Test

School	Teacher	Date
	DC	
St. Coletta of Washington DC	David Knight	11/15 & 11/22
Barnard Elementary School	Bill Piser	11/17
Columbia Heights Educational Campus	Sheila DeTorres	11/21
	VA	
Oakton Elementary School	Mary Marcantuono	11/15
Freedom Hill Elementary School	Elizabeth Dodd	11/15
Annandale High School	Christine Passut & Melissa Ainsworth	11/16
Oak View Elementary School	Cynthia Terry	11/17
Poplar Tree Elementary School	Amanda Moore	11/18
	MD	
Farmland Elementary School	Tatiana Khokhlova & Linda Bernard	11/22

- WIDA/SEA Forms Review - December 15, 2011:

  Following the copy editor's review of materials, a Forms Review was conducted at CAL with representatives from State Education Agencies (SEA) and specialists from the special education field in December 2011. Participants in the Forms Review examined all the test forms to ensure that the final materials were consistent and used this information to further refine the test forms. The participants in the Forms Review are listed in Table 2D.

Table 2D
Forms Review Participants

Name	Affiliation
Elizabeth Cranley	WIDA
Sandra Berndt	WIDA Consultant
Jennifer Christenson	CAL
Ann Evers	CAL
Deepak Ebenezer	CAL
Tatyana Vdovina	CAL
Cristina Sanchez-Lopez	Illinois Resource Center
Robert Fugate	Virginia Department of Education
Lia Mason	Virginia Department of Education
Sharon Prestridge	Mississippi Department of Education
Cindra Visser	Outside Consultant
Laurene Christensen	Minnesota Department of Education
Nancy Mullins	Maine Department of Education
Robert Romano	New Mexico Department of Education
Gaye Fedorchak	New Hampshire Department of Education
Carolyn Rosenberg	Maryland Department of Education

- Operational Field Test - March - June 2012:
  An operational field test of Alternate Access Series 100 took place from March to June 2012 in 15 states. (See Table 4A for a list of participating states.)

- Standard Setting Study – October 9-10, 2012
  A Standard Setting Study was conducted by experienced educators to establish rating scales and cut scores for assigning student scores to proficiency levels. The results of the Standard Setting Workshop are described in the *Alternate ACCESS for ELLs™ Standard Setting Study: Technical Brief* (CAL, 2012a).

- Psychometric Analysis – August - October, 2012
  Psychometric analyses were conducted at CAL to support development activities. The results of these analyses are presented in Chapter 7.

3 Administration of Alternate ACCESS™ Operational Field Test Series 100

Field testing of Alternate ACCESS was conducted from March 12 to June 1, 2012. In total, 1,912 students in grades 1-12 in 15 WIDA states participated in the field test. Participating SEAs

encouraged educators in their states to sign up for the field test through the regular ACCESS test ordering site provided by MetriTech, Inc.

The administrations were labeled as an operational field test, meaning states had the option of designating participation in the testing as a field test activity or as the first operational testing opportunity of the Alternate ACCESS program. States also had the option to decline participation in the operational field test.

4 Description of Operational Field Test Participants

A total of 1,912 students participated in the operational field test of Alternate ACCESS. Within the 15 participating states, students were drawn from 182 school districts. Data were cleaned to remove records for students who had been assigned to the incorrect grade-level cluster form, who did not have grade recorded, or who had been identified as not meeting the participation criteria. This resulted in 36 student records being removed from the dataset. All summaries and analyses included in this brief are based on the remaining 1,876 student records. In some cases, the totals in the following tables may not add to 100% due to rounding.

Table 4A shows the number of students who are included in the analyses presented by state, after the data were cleaned.

Table 4A

Students in the Operational Field Test by State

State	No. of students	Percent of students
DC	77	4%
DE	38	2%
KY	99	5%
MD	8	0%
ME	28	1%
MN	1	0%
MS	18	1%
ND	3	0%
NJ	19	1%
NM	299	16%
OK	216	12%
SD	56	3%
VA	985	53%
VT	10	1%
WI	19	1%
Total	1,876	100%

Demographic information for the participating students is shown in Table 4B through Table 4E.

Table 4B shows gender of the participating students. Of the 1,876 students who participated, 63% were boys (1,188 students) and 35% were girls (665 students). Gender information for 23 students was missing.

Table 4B

Students in the Operational Field Test by Gender

Gender	No. of students	Percent of students
Female	665	35%
Male	1,188	63%
Not Specified	23	1%
Total	1,876	100%

Table 4C shows the number of participating students by grade-level cluster. Students were approximately evenly spread across the grade-level clusters, with the greatest number of participants in grades 3-5 (32%) and the lowest number of participants in grades 1-2 (21%).

Table 4C

Students in the Operational Field Test by Grade Level Cluster

Grade-level cluster	No. of students	Percent of students
1-2	390	21%
3-5	599	32%
6-8	435	23%
9-12	452	24%
Total	1,876	100%

Table 4D shows the number of students by ethnicity; Table 4E shows the number of students by race. Over half of participating students were Hispanic (54%). The largest racial groups were Asian (22%) and White (31%). Many students (19%) had no race specified.

Table 4D

Students in the Operational Field Test by Ethnicity

Ethnicity	No. of students	Percent of students
Hispanic	1,017	54%
Non-Hispanic	772	41%
Not Specified	87	5%
Total	1,876	100%

Table 4E

Students in the Operational Field Test by Race

Race	No. of students	Percent of students
American Indian/Alaskan	365	19%
Asian	405	22%
Black	172	9%
Pacific Islander	12	1%
White	573	31%
Not Specified	349	19%
Total	1,876	100%

5 Item Analysis and Scaling

5.1 Overview

Because the focus of the operational field test was on the development of the new assessment, field test data were scored and analyzed to examine the psychometric quality of the Alternate ACCESS tasks. Since the assessment tasks and approach to administration were so new, the results presented in this section focus on how the task functioned. As with ACCESS, the Rasch measurement model was used to evaluate the psychometric quality of the tasks and the extent to which the tasks combine together to form a coherent test within a domain.

Scaling is the process of developing a standard reporting scale to make the scores on a test more usable to educators. The scaling procedure used for Alternate ACCESS involved putting student raw scores from all the test forms onto the same scale, so that results could be compared. The following sections detail how tasks were scored to produce raw scores and how Rasch procedures were used to create a reporting scale.

5.2 Assigning Raw Scores to Test Tasks

Students' original responses for all domains were converted to raw scores for psychometric analysis. For Listening and Reading, when students were unable to complete a task successfully, the prompt was repeated with increasing levels of support, allowing students multiple opportunities to respond to the selected response tasks. The repeated prompts are labeled as: CUE A: Initial Prompt, CUE B: Simplified Prompt, and CUE C: Simplified Prompt & Answer. The raw scores for these repeated prompts for Listening and Reading tasks are presented in Table 5A.

Table 5A

Raw Scores: Listening and Reading

Response category	Listening and Reading: All tasks (1-9)
Not Administered	0
No Response	0
Blank	0
Incorrect	1
Cue C	2
Cue B	3
Cue A	4

The Speaking section is a series of constructed response questions that include multiple opportunities within each task for a student to provide an answer. Unlike the domains of Listening and Reading, the levels of support remain constant during the multiple attempts. Student responses for the Speaking section were scored out of a possible two points (‘Approaches’ = 1, ‘Meets’ = 2) for each task.

The Writing section, comprising two parts, is a series of constructed response questions, each of which provides several opportunities for the student to produce an expected answer to a particular task. Similar to the Speaking section, the level of support on these tasks remains constant, but multiple opportunities are provided to allow students to perform at the expected proficiency level if they are able. The tasks in Writing Part A and B (Tasks 1-8) are highly constrained, and student responses are scored out of a possible two points (‘Approaches’ = 1, ‘Meets’ = 2) for each task. In Writing Part C (Tasks 9-10), test takers are provided with two open-ended tasks that are scored on a rubric related to the ACCESS proficiency levels. Examinee responses at level P1 of the rubric are recorded by the test administrator as ‘Meets 1,’ level P2 is recorded as ‘Meets 2’ and level P3 is recorded as ‘Meets 3.’ Thus, responses on each of these two tasks are scored out of a possible four points (‘No Response’ = 0, ‘Approaches’ = 1, ‘Meets 1’ = 2, ‘Meets 2’ = 3, and ‘Meets 3’ = 4)**.**

The raw score tables for the Speaking and Writing tasks are presented in Table 5B. Writing is divided into Parts A and B, and Part C.

Table 5B

Raw Scores : Speaking and Writing

Response category	Speaking: All tasks (1-8)	Writing: Parts A & B (1-8)	Writing: Part C (9-10)
Not Administered	0	0	0
No Response	0	0	0
Blank	0	0	0
Approaches	1	1	1
Meets	2	2	N/A
Meets 1	N/A	N/A	2
Meets 2	N/A	N/A	3
Meets 3	N/A	N/A	4

5.2.1 Raw Scores Statistics

This section presents the results, in terms of raw scores, of student performances on the operational field test of Alternate ACCESS. Tables 5C to 5F present the minimum score, maximum score, mean, and standard deviation of the raw scores by grade-level cluster and domain.

Table 5C

Raw Score Descriptive Statistics: Listening

Grade-level cluster	No. of students	Min	Max	Mean	Std. Dev.
1-2	387	0	36	26.89	11.21
3-5	595	0	36	30.09	9.45
6-8	431	0	36	30.50	9.43
9-12	446	0	36	28.83	10.48

Table 5D

Raw Score Descriptive Statistics: Reading

Grade-level cluster	No. of students	Min	Max	Mean	Std. Dev.
1-2	388	0	36	25.29	11.27
3-5	593	0	36	28.81	10.34
6-8	433	0	36	28.67	9.76
9-12	442	0	36	27.46	10.32

Table 5E

Raw Score Descriptive Statistics: Speaking

Grade-level cluster	No. of students	Min	Max	Mean	Std. Dev.
1-2	385	0	16	11.31	5.92
3-5	591	0	16	13.09	4.94
6-8	426	0	16	12.94	5.20
9-12	435	0	16	11.94	5.76

Table 5F

Raw Score Descriptive Statistics: Writing

Grade-level cluster	No. of students	Min	Max	Mean	Std. Dev.
1-2	367	0	24	12.35	6.52
3-5	574	0	24	15.57	6.69
6-8	407	0	24	16.95	6.42
9-12	426	0	24	15.12	7.31

5.3 Item Analysis

For the item analysis, Rasch measurement approaches as applied in the Winsteps software program (Linacre & Wright, 2006) were used. The measurement model that formed the basis of the Alternate ACCESS analyses was the Rasch Rating Scale model (Andrich, 1978), as this model is appropriate for polytomously scored test tasks. The Rating Scale model specifies that all tasks in a test have the same number of response categories and that they share the same rating scale structure. In other words, all tasks in a test have the same step or threshold parameters. The Rasch Rating Scale model is appropriate for modeling Alternate ACCESS tasks because the same scoring rules are used to convert students' original responses to raw scores within each grade-level cluster and each domain. Therefore the test tasks in each grade-level cluster and domain share the same rating scale structure. That is to say, a "test" comprises the tasks on one domain in one grade-level cluster; for example, 6-8 Reading. Mathematically, the rating scale model can be represented as

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - F_k$$

where

P_{nik} = probability of person "n" on task "i" responding at level "k" on the rating scale

P_{nik-1} = probability of person "n" on task "i" responding at level "k - 1" on the rating scale (i.e., the next lowest rating)

B_n = ability of person "n"

D_i = difficulty of task "i"

F_k = calibration of step "k" on the rating scale

For the Writing tasks, two separate rating scales were used, one for each part of the assessment. The tasks in Writing Parts A and B (Tasks 1-8) were modeled using a rating scale with two categories and the tasks in Writing Part C (Tasks 9-10) were modeled using a rating scale with five categories, as described in Chapter 5.2.

All Rasch analyses were conducted using the Rasch measurement software program Winsteps (Linacre, 2006). Rasch statistics are presented in several of the tables that follow. When referring to the measure of examinee ability, the term "ability measure" is used throughout this report (rather than *theta*, which is used commonly when discussing models based on Item Response Theory [IRT]). When speaking of the measure of how difficult an item was, the term "item difficulty measure" is used (rather than the term *b parameter*, used commonly when discussing models based on IRT). "Step measures" refers to the calibration of the steps in the Rasch Rating Scale model presented above. All three measures (ability, difficulty, and step) are expressed in terms of Rasch logits, which then are converted into scores on the Alternate ACCESS for ELLs™ score scale for reporting purposes.

For the initial Rasch calibration, the Rasch analyses were conducted separately by grade-level cluster and domain; therefore, the parameters for each grade-level cluster and domain were expressed on a unique logit scale. In the later stages of the psychometric analysis (see Chapter 5.4), the step or threshold parameters were constrained to be equal across grade-level clusters by domain through an anchoring process in order to put the task parameters across grade-level clusters by domain on the same logit scale.

After the initial Rasch calibration of the grade-level cluster, the fit statistics of the tasks were examined to ensure that all tasks were contributing to the measurement of the target construct. Fit statistics help to identify items or persons that do not meet the Rasch specifications; in other words, they pinpoint those items or persons that are extreme in terms of model misfit and therefore may distort the interpretation of results. Items or persons that do not meet the Rasch measurement model specifications are not automatically rejected, but are examined to identify in what way, and why, they fall short of the model specifications, and whether, on balance, they contribute to or corrupt measurement (Smith, 1996).

Since this is the first operational field test administration of a new assessment developed for a very unique student population, a somewhat liberal approach was taken in terms of the criteria used for identifying potentially problematic tasks. These criteria were used only to flag tasks that need further examination rather than rejecting tasks.

Criteria suggested by Smith (1996) for examining the fit of polytomous tasks to the Polytomous Rasch model were used. Smith suggested that mean-square outfit statistics above 1.50 indicate that something other than the construct being measured is influencing responses to the items. Thus, a task with a high fit statistic has the potential of distorting or degrading the measurement system and therefore is an important concern (Linacre, 2002), while a task with a low fit statistic is less productive for measurement, but not degrading and therefore not as big a concern.

Following the suggestions by Linacre (2002), large outfit statistics were examined before large infit statistics. Outfit mean-square statistics are influenced by outliers and can be skewed if students with extreme high-level or low-level proficiency do not perform as expected. In our analysis, tasks with an outfit mean-square statistic above 1.50 were flagged for internal examination by CAL. Next, tasks with large infit statistics were examined. High infit mean-squares indicate that the tasks are misperforming for the population for whom the tasks are targeted. This is a bigger threat to validity, but is more difficult to diagnose than high outfit (Linacre, 2002). Because infit statistics may be inflated when a few respondents have unusual response patterns on tasks that were designed to be very easy (for example, A1 tasks), tasks were only flagged for content review if the percentage of maximum possible score points was less than 85% and the infit mean-square statistic was greater than 1.50. The percent of maximum possible score points (PMPS) is obtained by dividing the average score by the maximum possible score point for that task, then multiplying by 100. It is basically a rescaling of the average score. The percentage of maximum possible score points is a common measure used to indicate the task difficulty for a polytomously scored task, with a higher value indicating an easier task. Tasks that were flagged using these infit criteria were subsequently examined by a Content review panel. This was done because greater attention should be paid to the infit statistics, rather than the outfit statistics, as the infit statistics have been adjusted for outliers and are therefore less susceptible to inflation from inconsistent or unexpected responses.

5.3.1 Item Statistics

This section presents the results, in terms of classical item statistics and Rasch statistics, of student performances on the operational field test of Alternate ACCESS. The following tables, 5G to 5V, present the results of the field test analyses on the items by their respective domains. In each table, the first column presents the task name. Within the task name, the first letter indicates the domain (L for Listening, R for Reading, S for Speaking, and W for Writing) and the first number indicates the task number. The next letter and number indicate the proficiency level of the task. The following column, labeled "Count," shows the total number of students included in the analysis. The next column shows the average score, which is the total number of points awarded to all students divided by the number of students taking the task. The next column shows the PMPS, explained above. It was computed by rescaling the average score to a value of 0 to 100 percent. The following column, labeled “Measure,” is the logit measure that is produced by the Rasch calibration. The next column, “Error,” shows the statistical precision with which

the measure value has been calculated. As with any statistic, the more data there are to estimate it (i.e., the larger the number in the "Count" column), the smaller the error will be.

The next two columns show the infit and outfit mean-square statistics that were used to identify items that were not fitting the Rasch measurement model, as discussed above.

The last column is labeled "PTPS," which stands for a point polyserial correlation for polytomously scored items. As a correlation between a student's performance on an item and the student's measure, its range extends from -1.00 to +1.00. For well-functioning items, this correlation must be positive, as it indicates an item's discrimination (i.e., how well it separates lower scorers from higher scorers). As an indication of discrimination, for measures developed using the Rasch model, it is expected that these correlation coefficients should generally fall within a common range for any single measure or analysis. Thus, this coefficient will tend to vary out of a common range with items that also show "misfit" through the infit and outfit mean-square statistics. Even when fit to the Rasch measurement model is not a problem, this correlation can also be low when an item is extremely easy (i.e., almost everyone is getting it correct) or extremely difficult (i.e., almost everyone is getting it incorrect). Although such items can fit the Rasch measurement model well, they cannot effectively discriminate between lower and higher ability students.

5.3.1.1 Listening

Table 5G shows the item statistics for Listening 1-2. In terms of outfit mean-square, only one task, Task L1_A1, was above the threshold of 1.50. Task L1_A1 is the easiest task on the test, with an average score of 3.62 or PMPS of 91%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses; for example if students with high-level proficiency do not perform as well as expected, this can impact the outfit mean-square statistic.

Table 5G
Listening 1-2 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
L1_A1	387	3.62	91%	-2.15	0.12	2.91	8.46	0.53
L2_A2	387	2.57	64%	0.86	0.06	1.06	1.08	0.74
L3_A2	387	3.01	75%	0.18	0.07	0.93	0.83	0.85
L4_A3	387	3.29	82%	-0.53	0.09	0.89	0.72	0.86
L5_A3	387	3.14	78%	-0.10	0.08	0.80	0.62	0.88
L6_P1	387	3.18	79%	-0.19	0.08	0.86	0.67	0.87
L7_P1	387	2.51	63%	0.94	0.06	0.89	0.81	0.77
L8_P2	387	3.06	77%	0.07	0.07	0.87	0.71	0.87
L9_P2	387	2.52	63%	0.93	0.06	0.96	0.83	0.77

Table 5H shows the item statistics for Listening 3-5. In terms of outfit mean-square, only one task, Task L1_A1, was above the threshold of 1.50. Task L1_A1 is the easiest task on the test, with an average score of 3.72 or PMPS of 93%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses.

Table 5H

Listening 3-5 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
L1_A1	595	3.72	93%	-1.61	0.12	3.92	3.91	0.59
L2_A2	595	3.51	88%	-0.31	0.08	1.42	0.97	0.82
L3_A2	595	3.54	89%	-0.45	0.09	1.06	0.76	0.85
L4_A3	595	3.53	88%	-0.43	0.09	1.08	0.65	0.86
L5_A3	595	3.18	80%	0.60	0.06	1.01	0.79	0.80
L6_P1	595	3.17	79%	0.64	0.06	0.93	0.72	0.81
L7_P1	595	3.51	88%	-0.33	0.08	0.99	0.42	0.88
L8_P2	595	3.26	82%	0.44	0.06	0.92	0.74	0.84
L9_P2	595	2.68	67%	1.46	0.05	0.99	0.99	0.70

Table 5I shows the item statistics for Listening 6-8. In terms of outfit mean-square, only one task, Task L1_A1, was above the threshold of 1.50. Task L1_A1 is the easiest task on the test, with an average score of 3.82 or PMPS of 96%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses.

Table 5I

Listening 6-8 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
L1_A1	431	3.82	96%	-2.36	0.15	2.00	3.39	0.55
L2_A2	431	3.39	85%	0.17	0.08	1.18	0.99	0.82
L3_A2	431	3.43	86%	0.05	0.09	0.86	0.81	0.87
L4_A3	431	3.46	87%	-0.05	0.09	1.20	1.10	0.83
L5_A3	431	3.38	85%	0.21	0.08	0.88	0.67	0.86
L6_P1	431	3.05	76%	0.95	0.07	1.10	1.00	0.78
L7_P1	431	3.40	85%	0.15	0.08	0.84	0.59	0.88
L8_P2	431	3.37	84%	0.22	0.08	1.12	1.02	0.83
L9_P2	431	3.19	80%	0.67	0.07	0.87	0.71	0.84

Table 5J shows the item statistics for Listening 9-12. In terms of outfit mean-square, only one task, Task L1_A1, was above the threshold of 1.50. Task L1_A1 is the easiest task on the test,

with an average score of 3.73 or PMPS of 93%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses.

Table 5J

Listening 9-12 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
L1_A1	446	3.73	93%	-2.33	0.14	2.33	9.90	0.56
L2_A2	446	3.25	81%	0.12	0.08	1.45	1.40	0.80
L3_A2	446	3.41	85%	-0.43	0.09	1.11	0.83	0.86
L4_A3	446	2.97	74%	0.72	0.06	0.96	1.00	0.80
L5_A3	446	3.36	84%	-0.22	0.09	0.67	0.44	0.91
L6_P1	446	2.95	74%	0.74	0.06	0.88	0.78	0.83
L7_P1	446	3.36	84%	-0.24	0.09	0.86	0.49	0.89
L8_P2	446	2.78	70%	1.02	0.06	0.94	0.83	0.79
L9_P2	446	3.02	76%	0.62	0.07	0.95	0.76	0.84

Overall, the fit statistic analyses indicate that the Listening tasks in all four grade clusters fit the Rasch measurement model appropriately with the exception of the first task, which exhibited high outfit mean-square statistics. These tasks were designed to measure the lowest level of proficiency, A1, so they tend to be extremely easy. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses. Several qualities of the A1 tasks make them different from other tasks that appear on Alternate ACCESS in terms of their characteristics and demands. For the Listening domain, the task level demand for A1 Alternate Model Performance Indicators across all grade clusters was to "Attend" to oral instruction (See AMPIs at http://wida.us/assessment/alternateaccess.aspx). Hence, A1 tasks across all grade clusters for Listening were designed to measure whether a student can pay attention to the test administrators. Test administrators scoring this A1 Task were only to judge whether a student had paid attention to them during the reading of the Test Administrator Script. Additionally, a student could have up to four opportunities (CUE A – 2, CUE B – 1, CUE C – 1) hearing the Test Administrator reading through the task to pay attention. Unlike other tasks in the Listening domain, the student was not asked directly to select the correct answer from the three possible options. The excerpt below from the *Test Administrator Manual* (WIDA, 2012a) describes the scoring criteria for A1 tasks for Listening:

> Please note that for Task 1, where the student is required to "attend" or "acknowledge," the test administrator should rate the student's response as correct if there is evidence that the student is engaged in the test task by paying attention. The evidence of engagement through attention on the part of the student can vary and may manifest itself in a variety of ways. For example, a student may demonstrate his or her engagement by looking at the response option, by nodding, by placing an object on the correct response option, etc. In order to allow the student to demonstrate his or

her proficiency, any evidence of engagement that is typical for that student in an instructional setting should be rated as a correct response (p.41).

5.3.1.2 Reading

Table 5K shows the item statistics for Reading 1-2. In terms of outfit mean-square, only one task, Task R1_A1, was above the threshold of 1.50. Task R1_A1 is the easiest task on the test, with an average score of 3.70 or PMPS of 93%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses; for example, if students with high-level proficiency do not perform as well as expected, this statistic can be impacted.

Table 5K

Reading 1-2 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
R1_A1	388	3.70	93%	-3.18	0.15	2.34	6.42	0.49
R2_A2	388	3.16	79%	-0.62	0.09	1.43	1.14	0.81
R3_A2	388	3.19	80%	-0.70	0.09	1.09	0.75	0.84
R4_A3	388	3.10	78%	-0.43	0.08	1.21	1.07	0.84
R5_A3	388	2.60	65%	0.70	0.07	1.36	1.20	0.80
R6_P1	388	2.55	64%	0.79	0.07	1.06	0.95	0.83
R7_P1	388	2.45	61%	0.97	0.07	0.68	0.67	0.87
R8_P2	388	2.07	52%	1.57	0.06	0.71	0.71	0.80
R9_P2	388	2.48	62%	0.91	0.07	0.87	0.64	0.85

Table 5L shows the item statistics for Reading 3-5. In terms of outfit mean-square, only one task, Task R1_A1, was above the threshold of 1.50. Task R1_A1 is the easiest task on the test, with an average score of 3.78 or PMPS of 95%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses.

Table 5L

Reading 3-5 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
R1_A1	593	3.78	95%	-3.11	0.14	2.51	6.65	0.53
R2_A2	593	3.40	85%	-0.44	0.09	1.70	1.35	0.82
R3_A2	593	3.48	87%	-0.82	0.09	1.13	0.96	0.84
R4_A3	593	3.27	82%	0.07	0.08	1.20	1.15	0.86
R5_A3	593	3.26	82%	0.11	0.08	0.99	0.73	0.89
R6_P1	593	2.87	72%	1.14	0.06	0.85	0.93	0.86
R7_P1	593	3.08	77%	0.64	0.07	0.90	0.73	0.88
R8_P2	593	2.67	67%	1.55	0.06	0.71	0.66	0.82
R9_P2	593	3.00	75%	0.85	0.06	0.97	0.72	0.86

Table 5M shows the item statistics for Reading 6-8. In terms of outfit mean-square, only one task, Task R1_A1, was above the threshold of 1.50. Task R1_A1 is the easiest task on the test, with an average score of 3.81 or PMPS of 95%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses.

Table 5M

Reading 6-8 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
R1_A1	433	3.81	95%	-3.78	0.19	2.58	4.47	0.54
R2_A2	433	3.49	87%	-0.79	0.12	1.42	1.06	0.85
R3_A2	433	3.52	88%	-0.98	0.12	1.55	0.96	0.84
R4_A3	433	3.31	83%	0.10	0.10	1.27	0.92	0.87
R5_A3	433	3.49	87%	-0.75	0.12	1.27	0.81	0.85
R6_P1	433	2.91	73%	1.27	0.07	0.97	0.84	0.82
R7_P1	433	2.95	74%	1.19	0.07	0.87	0.84	0.85
R8_P2	433	2.64	66%	1.79	0.06	0.86	0.92	0.79
R9_P2	433	2.55	64%	1.95	0.06	0.95	0.87	0.75

Table 5N shows the item statistics for Reading 9-12. In terms of outfit mean-square, only one task, Task R1_A1, was above the threshold of 1.50. Task R1_A1 is the easiest task on the test, with an average score of 3.71 or PMPS of 93%. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses.

Table 5N

Reading 9-12 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
R1_A1	442	3.71	93%	-2.91	0.16	3.03	8.05	0.59
R2_A2	442	3.42	86%	-0.78	0.1	1.25	0.98	0.84
R3_A2	442	3.45	86%	-0.91	0.11	1.03	0.48	0.86
R4_A3	442	3.20	80%	0.03	0.08	1.27	1.06	0.84
R5_A3	442	3.44	86%	-0.90	0.11	1.33	0.84	0.82
R6_P1	442	2.62	66%	1.27	0.06	0.99	0.87	0.78
R7_P1	442	2.81	70%	0.94	0.07	0.90	1.02	0.84
R8_P2	442	2.46	62%	1.54	0.06	0.85	0.79	0.77
R9_P2	442	2.35	59%	1.71	0.06	0.83	0.78	0.75

Overall, the fit statistic analyses indicate that the Reading tasks in all four grade clusters fit the Rasch measurement model appropriately with the exception of the first task in all four grade clusters, which showed high outfit mean-square statistics. Similar to the first task of the tests for the Listening domain, these tasks were also designed to measure the lowest level of proficiency, A1, so they tend to be extremely easy. The outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses. Several qualities of the A1 tasks make them different from other tasks that appear on Alternate ACCESS in terms of their characteristics and demands. For the Reading domain, the task level demand for A1 AMPIs across all grade clusters was to "Attend" to presented text or graphics. (See a description of the Alternate Model Performance Indicators on the WIDA website at http://wida.us/assessment/alternateaccess.aspx.) Hence, A1 tasks across all grade clusters for Reading were designed to measure whether a student can pay attention to the text or images placed in front of them by a test administrator. Test administrators scoring this A1 Task were only to judge whether a student had paid attention to them during the presentation of the text or graphics in the Test Booklet. Additionally, a student had to four opportunities (CUE A – 2, CUE B – 1, CUE C – 1) to pay *attention* to the text or graphic presented in the Test Booklet. Unlike other tasks in the Reading domain, the student was not asked directly to select the correct answer from the three possible options. The excerpt below from the *Test Administrator Manual* describes the scoring criteria for A1 tasks for Reading:

> Please note that for Task 1, where the student is required to "attend" or "acknowledge," the test administrator should rate the student's response as correct if there is evidence that the student is engaged in the test task by paying attention. The evidence of engagement through attention on the part of the student can vary and may manifest itself in a variety of ways. For example, students may demonstrate their engagement by looking at the response option, by nodding, by placing an object on the correct response option, etc. In order to allow the student to demonstrate his or her proficiency, any evidence of engagement that is typical for that student in an instructional setting should be rated as a correct response. (p. 46)

5.3.1.3 Speaking

Table 5O shows the item statistics for Speaking 1-2. In terms of outfit mean-square, only one task, Task S1_A1, was above the threshold of 1.50. Task S1_A1 is the easiest task on the test, with an average score of 1.61 or PMPS of 81%.

In terms of infit mean-square, Task S1_A1 also had an infit mean-square statistic above the threshold of 1.50. Furthermore, since the task's PMPS (81%) is lower than the prescribed cutoff of 85% (see Chapter 5.3), this task was flagged for content review. The content experts from CAL and WIDA reviewed the content of this task and determined that there was no discernible content concern. The AMPI language demand for this task is that the student *vocalizes* to communicate about everyday objects. This task asks the student to vocalize in response to the question, "Can you say paper?" The content of this question appears to be measuring the targeted AMPIs appropriately and thus is contributing to the measurement of the target construct.

Table 5O

Speaking 1-2 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
S1_A1	385	1.61	81%	-2.66	0.27	1.71	1.98	0.86
S2_A2	385	1.56	78%	-1.59	0.22	1.39	0.80	0.89
S3_A3	385	1.49	75%	-0.41	0.19	1.24	0.71	0.91
S4_A1	385	1.54	77%	-1.21	0.21	0.99	1.23	0.92
S5_A2	385	1.45	73%	0.01	0.18	1.23	1.22	0.90
S6_A3	385	1.47	74%	-0.23	0.19	1.10	0.51	0.91
S7_P1	385	1.31	66%	1.52	0.15	0.89	0.57	0.86
S8_P2	385	0.89	45%	4.57	0.13	0.52	0.72	0.68

Table 5P shows the item statistics for Speaking 3-5. In terms of outfit mean-square, two tasks, Task S1_A1 and S2_A2, were above the threshold of 1.50. These are two of the easiest tasks on the test.

Table 5P
Speaking 3-5 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
S1_A1	591	1.76	88%	-2.15	0.24	1.79	2.91	0.85
S2_A2	591	1.74	87%	-1.31	0.21	1.42	1.55	0.88
S3_A3	591	1.71	86%	-0.63	0.18	1.18	0.98	0.90
S4_A1	591	1.72	86%	-0.84	0.19	0.93	0.40	0.92
S5_A2	591	1.68	84%	-0.08	0.17	1.14	1.03	0.90
S6_A3	591	1.69	85%	-0.38	0.18	0.78	0.57	0.92
S7_P1	591	1.58	79%	1.16	0.14	0.99	0.73	0.86
S8_P2	591	1.21	61%	4.23	0.11	0.72	1.00	0.64

Table 5Q shows the item statistics for Speaking 6-8. For the outfit mean-square statistics, none of the tasks was above the threshold of 1.50.

Table 5Q
Speaking 6-8 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
S1_A1	426	1.73	87%	-1.99	0.28	1.50	0.80	0.89
S2_A2	426	1.69	85%	-0.89	0.23	1.33	1.08	0.91
S3_A3	426	1.67	84%	-0.54	0.22	1.09	0.65	0.92
S4_A1	426	1.69	85%	-1.00	0.24	0.85	0.52	0.93
S5_A2	426	1.68	84%	-0.73	0.23	0.83	0.85	0.93
S6_A3	426	1.67	84%	-0.49	0.22	0.77	0.84	0.93
S7_P1	426	1.53	77%	1.52	0.16	1.22	1.02	0.84
S8_P2	426	1.26	63%	4.12	0.15	0.85	1.39	0.69

Table 5R shows the item statistics for Speaking 9-12. In terms of outfit mean-square, only one task, Task S1_A1, was above the threshold of 1.50. Task S1_A1 is the easiest task on the test, with an average score of 1.66 or PMPS of 83%.

In terms of infit mean-square, Task S1_A1 also had an infit mean-square statistic above the threshold of 1.50. Moreover, since the task's PMPS (83%) is lower than the prescribed cutoff of 85%, this task was flagged for content review. The content experts from CAL and WIDA reviewed the content of this task and determined that there was no discernible content concern. The AMPI language demand for this task is that the student *vocalizes* in response to a question about everyday objects. This task asks the student to vocalize in response to the question, "Can

you say door?" The content of this question appears to be measuring the targeted AMPIs appropriately and thus is contributing to the measurement of the target construct.

Table 5R

Speaking 9-12 Item Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
S1_A1	435	1.66	83%	-2.95	0.30	1.66	1.82	0.87
S2_A2	435	1.58	79%	-0.84	0.20	1.22	0.85	0.91
S3_A3	435	1.56	78%	-0.45	0.19	1.38	0.75	0.90
S4_A1	435	1.57	79%	-0.72	0.20	1.15	0.95	0.91
S5_A2	435	1.56	78%	-0.49	0.19	0.83	0.51	0.93
S6_A3	435	1.54	77%	-0.17	0.18	0.72	0.53	0.94
S7_P1	435	1.36	68%	1.81	0.14	0.93	0.76	0.85
S8_P2	435	1.11	56%	3.81	0.13	0.98	1.43	0.71

Overall, the fit statistic analyses indicate that the Speaking tasks in all four grade clusters fit the Rasch measurement model appropriately with the exception that the first task for three out of the four grade cluster tests. These tasks showed high outfit mean-square statistics and two of them were also flagged for having high infit mean-square statistics.

For the tasks that were flagged for high outfit statistics, CAL reviewed the tasks but did not have discernible content concerns. The language demands for A1 Speaking tasks across grade cluster are such that *any vocalization* from the student during the task question is a demonstration of having met the expectation of the task. Additionally, the student has the possibility of hearing the task questions six times in total (refer to Chapter 1.5). The Test Administration Manual describes the expectation for the A1 Speaking task as:

> **Expectations at Level A1: Initiating.** To score 'Meets' on a task at Level A1 students must be able to **vocalize** in response to a question (e.g., incoherent but communicative vocalizations, such as grunts). There is no requirement at this level for students to use complete words.

For the two tasks that were also flagged for having high infit mean-square, the content experts from CAL and WIDA reviewed the tasks but did not have discernible content concerns. These tasks appear to be measuring the targeted AMPIs appropriately and thus are contributing to the measurement of the target construct.

5.3.1.4 Writing

Table 5S shows the item statistics for Writing 1-2. In terms of outfit mean-square, five tasks, W1_A1, W2_A2, W3_A3, W4_P1, and W7_A3 were above the threshold of 1.50. Tasks W1_A1 and W2_A2 are two of the easiest tasks on the test. The outfit mean-square statistic, which is not weighted for ability, may be inflated for these tasks due to some random responses

(as discussed in the other domains). For Task W4_P1 and Task W7_A3, it was also observed that some students who the model predicts should have received a higher score did not. Task W3_A3 was flagged based on both the outfit and infit/PMPS criterion, and it is discussed below. In terms of infit, Task W3_A3 had an infit mean-square statistic above the threshold of 1.50 and the task's PMPS (70%) is lower than the designated cutoff of 85%. Therefore, this task was flagged for content review. The content experts from CAL and WIDA reviewed the content of this task and determined that there was no discernible content concern. The AMPI language demand for this task was to *copy a word.* This task asks the student to copy the word "rules" and thus appears to be measuring the targeted AMPIs appropriately and contributing to the measurement of the target construct.

Table 5S

Writing Item 1-2 Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
W1_A1	367	1.77	89%	-7.54	0.28	1.35	1.61	0.61
W2_A2	367	1.67	84%	-5.05	0.24	0.89	8.51	0.76
W3_A3	367	1.39	70%	-1.31	0.16	1.51	2.14	0.81
W4_P1	367	1.01	51%	1.98	0.15	1.28	2.84	0.83
W5_A1	367	1.52	76%	-2.69	0.18	1.15	0.50	0.79
W6_A2	367	1.48	74%	-2.21	0.17	0.97	0.47	0.81
W7_A3	367	1.28	64%	-0.28	0.15	0.63	2.01	0.87
W8_P1	367	0.90	45%	2.85	0.14	0.70	0.57	0.86
W9_P3	367	0.69	17%	6.98	0.11	0.86	0.52	0.64
W10_P3	367	0.63	16%	7.28	0.12	0.70	0.43	0.60

Table 5T shows the item statistics for Writing 3-5. In terms of outfit mean-square, three tasks, Tasks W1_A1, W2_A2, and W4_P1 were above the threshold of 1.50. Task W1_A1 and W2_A2 are two of the easiest tasks on the test, so the outfit mean-square statistic, which is not weighted for ability, may be inflated for this very easy task due to some unexpected responses; for example if students with high-level proficiency do not perform as well as expected, this statistic can be affected. For Task W4_P1, it was also observed that some students received lower scores than the model predicted.

Table 5T

Writing Item 3-5 Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
W1_A1	574	1.83	92%	-6.66	0.32	1.45	9.90	0.63
W2_A2	574	1.78	89%	-4.34	0.24	1.30	9.90	0.74
W3_A3	574	1.63	82%	-1.34	0.15	1.18	0.69	0.83
W4_P1	574	1.37	69%	1.66	0.13	1.18	3.78	0.85
W5_A1	574	1.71	86%	-2.59	0.18	1.18	0.52	0.76
W6_A2	574	1.70	85%	-2.40	0.18	0.97	0.28	0.78
W7_A3	574	1.56	78%	-0.50	0.14	0.75	0.44	0.85
W8_P1	574	1.30	65%	2.33	0.13	0.85	1.00	0.87
W9_P3	574	1.32	33%	7.02	0.08	0.88	0.84	0.71
W10_P3	574	1.37	34%	6.84	0.08	0.79	0.74	0.67

Table 5U shows the item statistics for Writing 6-8. In terms of outfit mean-square, three tasks, W1_A1, W2_A2, and W3_A3, were above the threshold of 1.50. These tasks are among the easiest tasks on the test; thus, the outfit mean-square statistic, which is not weighted for ability, may be inflated for these very easy tasks due to some unexpected responses.

Table 5U

Writing Item 6-8 Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
W1_A1	407	1.87	94%	-4.83	0.30	1.61	2.84	0.61
W2_A2	407	1.83	92%	-3.43	0.25	1.01	9.90	0.73
W3_A3	407	1.71	86%	-1.23	0.19	1.49	8.01	0.79
W4_P1	407	1.50	75%	1.21	0.15	1.03	1.50	0.84
W5_A1	407	1.75	88%	-1.87	0.21	1.12	0.32	0.77
W6_A2	407	1.76	88%	-1.96	0.21	0.78	0.18	0.79
W7_A3	407	1.66	83%	-0.53	0.18	0.63	1.29	0.86
W8_P1	407	1.48	74%	1.40	0.15	0.77	0.64	0.87
W9_P3	407	1.64	41%	5.78	0.09	0.91	1.13	0.73
W10_P3	407	1.74	44%	5.46	0.09	0.90	1.32	0.67

Table 5V shows the item statistics for Writing 9-12. In terms of outfit mean-square, five tasks, W1_A1, W2_A2, W4_P1, W5_A1, and W8_P1, were above the threshold of 1.50. Tasks W1_A1 and W2_A2 are two of the easiest tasks on the test, so the outfit mean-square statistic, which is not weighted for ability, may be inflated for these easy tasks due to some unexpected responses. For Tasks W4_P1, W5_A1, and W8_P1, it was observed that some students received lower scores than the model predicted.

Table 5V

Writing Item 9-12 Statistics

Task name	Count	Average score	PMPS	Measure	Error	IN. MSQ	OUT. MSQ	PTPS
W1_A1	426	1.75	88%	-4.76	0.25	1.71	9.90	0.67
W2_A2	426	1.70	85%	-3.55	0.22	0.70	9.90	0.81
W3_A3	426	1.60	80%	-1.77	0.18	1.39	1.24	0.84
W4_P1	426	1.31	66%	1.36	0.14	1.21	3.43	0.84
W5_A1	426	1.60	80%	-1.84	0.18	0.95	2.30	0.83
W6_A2	426	1.59	80%	-1.64	0.18	0.73	0.32	0.85
W7_A3	426	1.52	76%	-0.77	0.17	0.54	0.21	0.89
W8_P1	426	1.30	65%	1.44	0.14	0.73	2.41	0.89
W9_P3	426	1.37	34%	5.78	0.09	0.80	0.79	0.73
W10_P3	426	1.37	34%	5.76	0.09	0.82	0.80	0.69

A few Writing tasks showed high outfit mean-square statistics, and one Writing task was flagged for a high infit mean-square value. The high outfit mean-square statistics were associated with student responses that were lower than expected. One particular task (W4_P1) was flagged as having high outfit mean-square statistics in three out of four grade clusters and therefore was investigated further by CAL. The content experts from CAL were asked to review these three tasks, and they were all determined to have met the appropriate AMPIs. Additionally, the percent correct index (PMPS) and Rasch measure show that for all grade clusters, the P1 tasks (W4 and W8), were always more challenging than the A1-A3 tasks in each folder. Therefore these P1 tasks, although flagged as having higher than usual outfit statistics, are measuring a higher level of language usage than the A1-A3 tasks, as designed.

As one task in Writing 1-2 was flagged for both high outfit and infit mean-square, the content experts from CAL and WIDA reviewed the tasks but did not find content concern. This task appears to be measuring the targeted AMPIs appropriately and thus is contributing to the measurement of the target construct.

5.4 Scaling

For Alternate ACCESS, a three-digit scale score (910 to 960) was selected to aid in score interpretation. The scale needed an interpretive center point across domains and composites, so the centering value of 935 was chosen to represent the midpoint of the cut score between proficiency levels A3 and P1 for the Grade 3-5 cluster. This is analogous to the ACCESS scale, where the score of 350 is set as the center value and represents the cut score between proficiency levels P3 and P4 for grade 5 (for more information see *Development and Field Test of ACCESS for ELLs® [WIDA Consortium Technical Report No. 1]*; Kenyon, 2006).

The procedure for developing the Alternate ACCESS scale was complex and involved a number of steps. These steps were carried out separately for each of the four domains until the last step,

when the separate domain scales were combined to form the composite scores. These steps are briefly summarized below.

5.4.1 Calibrating All Grade Level Tasks by Domain to a Common Scale

To facilitate the derivation of common cut scores (see *Alternate ACCESS for ELLs™ Standard Setting Study: Technical Brief*; CAL, 2012a), a common Rasch logit scale across grade-level clusters by domain was developed using the procedures described in this section.

As described in Chapter 5.3, after the initial Rasch calibration by grade-level cluster, task parameters for each grade-level cluster and domain were expressed on a unique logit scale. The most direct method to put parameters from different test forms on the same scale is to include some common tasks or common persons across forms in order to link the test forms across these tasks or persons. In the case of Alternate ACCESS, although there are no common tasks linking the tests across grade-level clusters, the constructs being measured and the specifications for the assessment tasks are the same across grade-level clusters within a given domain. In addition, the test blueprints across grade-level clusters by domain are the same, and the Alternate PLs and AMPIs for the test tasks across grade-level clusters pose nearly identical linguistic challenges and differ only in the topics presented. Lastly, the same scoring rules are used to convert students' original responses to raw scores by domain. Therefore, it is reasonable to model a single rating scale across all grade-level clusters by domain. This can be achieved by imposing the same threshold parameters across the four grade-level clusters by domain.

The grade 3-5 step or threshold parameters were used as the common step values, primarily because more grade 3-5 students participated in the field test, therefore producing more stable parameters than other grade-level clusters. For each domain, the grades 1-2, 6-8, and 9-12 rating scale threshold parameters were anchored to the grade 3-5 domain values using Winsteps. The difficulty parameters for grades 1-2, 6-8, and 9-12 were unanchored and thus were calibrated in the runs. All task parameters including the difficulty and threshold parameters were thus placed on the same logit scale across grade-level clusters by domain through this process. The logit scales were then transformed to the common reporting scale described in Chapter 5.4.2.

5.4.1.1 Anchoring the Rating Scale

This section presents the results of anchoring the grades 1-2, 6-8, and 9-12 rating scale step parameters to the grade 3-5 domain values using Winsteps. The characteristics of the rating scale were first examined to see whether the rating scale provided clearly defined and ordered categories and whether the raw score points supported the construction of a Rasch measure. For analytical purposes, the goal of this examination was to see whether the rating scale observations conformed reasonably closely to the specified model; when such conformity is not strictly met, possible explanations are sought out and possible remedies are explored (Linacre, 1999). Lastly, the original and anchored Rasch difficulty measures are reported.

The second column of each ‘Rating Scale statistics’ table shows the score category or score point. The next column, "Obs. count," shows the number of times each score occurred in the data, while "Obs. percent" shows the percentage of time each score was observed. The next column, “Obs. average measure,” shows the observed average measures of all responses receiving that score, while following column, "Expected measure," shows what the rating scale measurement model would predict as the measure corresponding to that score. The observed and expected measures should be close to one another if the rating scale fits the data. In general, observations in higher score categories must be produced by higher measures. This means that the observed average measure is expected to increase with category value. If the observed average measure does not increase with category value, it may indicate substantive problems with the rating scale category definition.

The next two columns show the step parameters. The "Original step measure" column reports the step parameters based on the separate grade-level cluster calibration while the "Final step measure" column reports the anchored values. The step parameter is the calibrated measure of the transition from the score below to the listed score, and it indicates how difficult it is to observe this score, not how difficult to obtain it. The step parameter is expected to increase with score value. Disordering of these estimates (such that they do not ascend in value up the rating scale), sometimes called "disordered deltas," indicates that the score is relatively rarely observed. In other words, the score occupies a narrow interval on the latent variable, and the score is never modal. However, disordering in step parameters does not introduce category disordering (as diagnosed by average measures) or category misfit (as diagnosed by mean-squares fit) (Linacre, 1999).

The last two columns are the infit and outfit mean-square statistics, which are the average of the respective mean-square statistics associated with the responses in each score point, meaning that they are sensitive to the frequency distribution of the score points. These statistics were based on the unanchored runs since anchoring can disturb the infit and outfit statistics. They show how well the data fit the Rasch rating scale. It is recommended that a well-functioning rating scale should have outfit mean-square statistics of less than 2.00 (Linacre, 2004). For a typical rating scale, a high outfit mean-square associated with a particular category may indicate that the category has been used by raters or respondents in unexpected ways. In the context of Alternate ACCESS, the score categories are simply raw scores converted from students’ original responses to the tasks. Therefore, a high outfit mean-square associated with a particular score point may indicate that the responses observed for that score point are unexpected and that there is more misinformation than information in the responses. Unexpected responses for an extreme score point, like the score point ‘0’ for Listening, is more likely to produce a high outfit mean-square than unexpected use of a central score such as ‘2.’ Also, since outfit mean-square statistics are sensitive to outliers, this misinformation may be confined to a few substantively explainable

responses. These two factors need to be brought into consideration when interpreting tasks that show high outfit mean-square statistics.

Each 'Original and Final Rasch Measure' table shows the original Rasch measure derived from separate grade-level cluster calibration followed by the final Rasch measure derived after anchoring to the 3-5 grade cluster rating scale.

5.4.1.2 Listening

Table 5W shows the rating scale statistics for Listening. Overall, the observed count and observed percent showed that score points '0' to '3' were observed much less frequently than score point '4' and that there is a dramatic increase in frequency from score point '3' to '4'. This pattern, though not optimal for obtaining stable rating scale parameters, reflects the nature of the test administration in two respects: a) students were given partial credit for trying to perform the tasks with differing degrees of support, and b) students were given credit for responding to CUE A, which means they may have had CUE A repeated to them. These numbers also showed that even though in this particular sample the majority of students seemed to be able to perform the tasks without much support, there were still some students in need of some support in order to demonstrate their proficiency in Listening.

In general, the observed average measure increased with category value and the observed average and expected measure were very close to each other for all score points. This indicates a good rating scale model.

There was some disordering of step parameters for score points '2' and '3' which indicates that these two score points were relatively rare. This is supported by their observed percent values, where these two score points had proportionally fewer responses than other score points and there was a big increase in the number of responses from score point '3' to '4'.

There were high outfit mean-square statistics (i.e., > 2.00) for score point '0' on grade clusters 1-2, 3-5, and 9-12, as well as for score point '2' on grade 1-2. Since these statistics aggregate fit information across score points and tasks, the detail fit statistics at the score point and task level were examined to try to identify potential sources of misfit.

Table 5W

Rating Scale Statistics for Listening

Cluster	Score point	Obs. count	Obs. percent	Obs. average measure	Expected measure	Original step measure	Final step measure	IN. MSQ	OUT. MSQ
1-2	0	498	14%	-1.87	-1.70	NONE	NONE	0.71	3.28
	1	310	9%	-0.07	-0.30	-0.83	-0.74	1.01	0.73
	2	159	5%	0.58	0.51	0.79	0.50	0.99	2.43
	3	286	8%	1.13	1.15	0.16	0.23	0.94	1.56
	4	2230	64%	1.93	1.94	-0.13	0.01	1.18	1.22
3-5	0	464	9%	-1.51	-1.39	NONE	NONE	0.99	3.32
	1	280	5%	0.04	-0.24	-0.74	-0.74	0.20	0.90
	2	204	4%	0.56	0.56	0.50	0.50	0.93	1.04
	3	410	8%	1.10	1.31	0.23	0.23	0.01	0.56
	4	3997	75%	2.20	2.18	0.01	0.01	0.31	1.23
6-8	0	310	8%	-1.75	-1.61	NONE	NONE	0.78	1.22
	1	191	5%	-0.06	-0.24	-0.73	-0.74	1.12	0.89
	2	129	3%	0.69	0.67	0.61	0.50	1.02	1.68
	3	301	8%	1.21	1.14	0.05	0.23	0.96	1.01
	4	2948	76%	2.04	2.06	0.07	0.01	1.17	1.10
9-12	0	434	11%	-1.88	-1.75	NONE	NONE	0.81	8.70
	1	276	7%	-0.08	-0.28	-0.89	-0.74	1.01	0.69
	2	159	4%	0.52	0.57	0.72	0.50	0.95	1.60
	3	314	8%	1.19	1.18	0.18	0.23	0.88	0.96
	4	2831	71%	2.03	2.04	0.00	0.01	1.23	1.34

In particular, the detailed score point frequencies and fit statistics by task were examined. Table 5X provides an example of such an investigation for Listening 3-5. The results show that the outfit mean-square for score point ‘0’ of Task 1 was extremely high (10.00) while the rest of the outfit mean-square statistics were all less than 3. Furthermore, the Task 1 score point frequencies showed a skewed distribution, where the majority of students (91%) received the maximum score on this task. This suggests that the parameter estimates for Task 1 may not be very stable. It was hypothesized that the extremely high outfit mean-square of score point ‘0’ of Task 1 may have impacted the overall fit of the rating scale. To investigate this hypothesis, this task was removed and the parameters and fit statistics were re-estimated. Subsequent outfit mean-square statistics for all score points were then less than 2 when this task was excluded from estimation. The same investigation was conducted for Listening 1-2 and 9-12 with similar findings. It was therefore concluded that the high outfit mean-square statistics for score points ‘0’ and ‘2’ on the Listening tests were due to some isolated cases of unexpected responses for those score points on Task 1. As discussed earlier, Task 1 of the Listening test is a very easy task and most students are expected to earn a high score on this task. Therefore unexpected responses on this task are very likely to produce high outfit mean-square statistics.

Overall, the fit statistics showed that the rating scale implemented clearly defined and ordered scores and that the rating scale is functioning reasonably well. Additionally, the raw score points support the construction of a Rasch measure.

Table 5X

Detailed Item Category Frequencies and Fit Statistics for Listening 3-5

Item	Data code	Score number	Data count	Data %	Average ability	S.E. mean	OUT. MSQ	PTBSE corr.	Score value
L1_A1	0	0	31	5	-2.89	0.40	10.00	-.53	0 NA/NR/Blank
	2	2	17	3	-0.19	0.34	3.30	-.23	2 Cue C
	3	3	8	1	0.71	0.44	1.00	-.08	3 Cue B
	4	4	539	91	2.41	0.06	2.70	.57	4 Cue A
L2_A2	0	0	44	7	-2.80	0.24	0.80	-.77	0 NA/NR/Blank
	1	1	18	3	0.08	0.26	2.00	-.18	1 Incorrect
	2	2	16	3	0.30	0.23	0.50	-.15	2 Cue C
	3	3	32	5	1.17	0.16	0.90	-.06	3 Cue B
	4	4	485	82	2.66	0.05	1.40	.70	4 Cue A
L3_A2	0	0	44	7	-2.81	0.23	0.80	-.78	0 NA/NR/Blank
	1	1	15	3	-0.49	0.33	1.60	-.25	1 Incorrect
	2	2	12	2	0.28	0.24	0.60	-.15	2 Cue C
	3	3	30	5	1.14	0.14	0.70	-.06	3 Cue B
	4	4	494	83	2.64	0.05	1.00	.74	4 Cue A
L4_A3	0	0	45	8	-2.86	0.23	1.10	-.79	0 NA/NR/Blank
	1	1	15	3	-0.02	0.23	2.00	-.20	1 Incorrect
	2	2	14	2	0.37	0.18	0.40	-.14	2 Cue C
	3	3	24	4	0.77	0.15	0.30	-.12	3 Cue B
	4	4	497	84	2.65	0.05	1.00	.77	4 Cue A
L5_A3	0	0	52	9	-2.51	0.23	0.40	-.83	0 NA/NR/Blank
	1	1	48	8	0.62	0.12	0.60	-.16	1 Incorrect
	2	2	35	6	1.06	0.12	0.60	-.06	2 Cue C
	3	3	64	11	1.66	0.07	0.80	.04	3 Cue B
	4	4	396	67	2.96	0.05	1.30	.59	4 Cue A
L6_P1	0	0	56	9	-2.39	0.23	0.50	-.85	0 NA/NR/Blank
	1	1	53	9	0.62	0.09	0.50	-.17	1 Incorrect
	2	2	30	5	1.35	0.10	0.70	.01	2 Cue C
	3	3	53	9	1.68	0.08	0.80	.05	3 Cue B
	4	4	403	68	2.94	0.05	1.10	.60	4 Cue A
L7_P1	0	0	58	10	-2.32	0.22	1.40	-.84	0 NA/NR/Blank
	1	1	12	2	0.01	0.15	0.70	-.17	1 Incorrect
	2	2	6	1	0.03	0.24	0.20	-.13	2 Cue C
	3	3	11	2	1.09	0.13	0.30	-.03	3 Cue B

	4	4	508	85	2.63	0.05	0.90	.82	4 Cue A
L8_P2	0	0	64	11	-2.12	0.22	0.50	-.86	0 NA/NR/Blank
	1	1	37	6	0.73	0.11	0.90	-.11	1 Incorrect
	2	2	12	2	0.88	0.21	0.50	-.06	2 Cue C
	3	3	51	9	1.67	0.09	0.80	.04	3 Cue B
	4	4	431	72	2.85	0.05	1.00	.65	4 Cue A
L9_P2	0	0	70	12	-1.91	0.22	0.30	-.87	0 NA/NR/Blank
	1	1	82	14	1.26	0.07	0.50	.00	1 Incorrect
	2	2	62	10	1.69	0.07	0.60	.08	2 Cue C
	3	3	137	23	2.25	0.06	1.50	.18	3 Cue B
	4	4	244	41	3.40	0.05	3.00	.37	4 Cue A

Table 5Y presents the Rasch difficulty measure before and after anchoring the rating scale threshold parameters of the 1-2, 6-8, and 9-12 grade clusters to the 3-5 grade cluster for Listening. The second column of each table shows the "Original difficulty measure" in logits, when they were independently calibrated (see Chapter 4). The third column, "Anchored difficulty measure" shows the same measures when placed on the 3-5 rating scale. Overall, the difference between the original and the anchored difficulty measure are fairly small, indicating that a single rating scale across all grade-level clusters is a reasonable approach to take.

Table 5Y

Listening Original and Anchored Rasch Difficulty Measures

Cluster	Task name	Original difficulty measure*	Anchored difficulty measure
1-2	L1_A1	-2.15	-2.16
	L2_A2	0.86	0.89
	L3_A2	0.18	0.17
	L4_A3	-0.53	-0.56
	L5_A3	-0.10	-0.12
	L6_P1	-0.19	-0.21
	L7_P1	0.94	0.98
	L8_P2	0.07	0.06
	L9_P2	0.93	0.96
6-8	L1_A1	-2.36	-2.37
	L2_A2	0.17	0.17
	L3_A2	0.05	0.05
	L4_A3	-0.05	-0.04
	L5_A3	0.21	0.21
	L6_P1	0.95	0.95
	L7_P1	0.15	0.15
	L8_P2	0.22	0.22
	L9_P2	0.67	0.67
9-12	L1_A1	-2.33	-2.28
	L2_A2	0.12	0.11
	L3_A2	-0.43	-0.44
	L4_A3	0.72	0.71
	L5_A3	-0.22	-0.23
	L6_P1	0.74	0.74
	L7_P1	-0.24	-0.25
	L8_P2	1.02	1.02
	L9_P2	0.62	0.61

*See Chapter 5.3

5.4.1.3 Reading

Table 5Z shows the rating scale statistics for Reading. Overall, the observed count and observed percent showed that score points '0' to '3' were observed much less frequently than score point '4;' in addition, there is a dramatic increase in frequency from score point '3' to '4'. While this pattern may not be optimal for obtaining stable rating scale parameters, it does reflect the nature of the test because: a) students were given partial credit for trying to perform the tasks with differing degrees of support, and b) students were given credit for responding to CUE A, even if CUE A had been repeated to them. These numbers also showed that although in this particular sample the majority of students seemed to be able to perform the tasks without a lot of support, there were still some students who required some support in order to demonstrate their proficiency in Reading.

In general, the observed average measure increased with category value and the observed average and expected measure were very close to each other at all score points. This indicates a good rating scale model.

There was some disordering of step parameters for score points '2' and '3,' which indicates that these two score points were relatively rare. This is supported by the observed percent values, as these two score points had proportionally fewer responses than other score points and there was a big increase from score point '3' to '4'.

There were high outfit mean-square statistics (i.e., > 2.00) for score point '0' on Reading 3-5 and 9-12 and for score point '2' on Reading 1-2, 3-5, and 6-8. As explained above (see 5.4.1.2), these outfit mean-square statistics aggregate fit information across score points and tasks; thus, the detailed fit statistics at the score point and task level were examined to try to identify potential sources of misfit. Investigations similar to those conducted for the Listening tests were also conducted for the Reading tests. It was concluded that these high outfit mean-square statistics were due to a few isolated cases of unexpected responses for score points '0' and '2' of Task 1. As discussed earlier, Task 1 of the Reading tests is a very easy task and most students are expected attain a high score on this task. Therefore, unexpected responses on this task are very likely to produce high outfit mean-square statistics.

Overall, the rating scale statistics showed that the rating scale implemented clearly defined and ordered scores and that rating scale is functioning reasonably well. In addition, the raw score points support the construction of a Rasch measure.

Table 5Z

Rating Scale Statistics for Reading

Cluster	Score point	Obs. count	Obs. percent	Obs. average measure	Expected measure	Original step measure	Final step measure	IN. MSQ	OUT. MSQ
1-2	0	597	17%	-2.27	-2.11	NONE	NONE	0.89	1.75
	1	299	9%	-0.31	-0.64	-0.87	-0.74	1.18	1.27
	2	182	5%	0.49	0.41	0.42	0.50	0.97	2.74
	3	505	14%	1.12	1.19	-0.26	0.23	0.84	1.26
	4	1909	55%	2.71	2.71	0.71	0.01	1.19	1.13
3-5	0	572	11%	-2.30	-2.14	NONE	NONE	0.79	3.88
	1	329	6%	-0.24	-0.51	-1.07	-1.07	1.17	1.24
	2	210	4%	0.60	0.46	0.46	0.46	1.06	2.33
	3	569	11%	1.19	1.29	0.13	0.13	0.85	0.96
	4	3657	69%	2.67	2.66	0.74	0.74	1.16	1.14
6-8	0	382	10%	-2.61	-2.60	NONE	NONE	0.89	1.94
	1	283	7%	-0.34	-0.39	-1.56	-1.07	1.13	1.31
	2	172	4%	0.76	0.49	0.60	0.46	1.04	2.12
	3	452	12%	1.15	1.43	0.07	0.13	1.10	1.04
	4	2608	67%	3.12	3.08	0.89	0.74	1.07	1.06
9-12	0	481	12%	-2.25	-2.20	NONE	NONE	0.93	7.86
	1	352	9%	-0.27	-0.41	-1.33	-1.07	1.11	1.09
	2	191	5%	0.50	0.38	0.65	0.46	0.97	1.09
	3	414	10%	1.19	1.52	0.11	0.13	1.06	0.99
	4	2540	64%	2.92	2.88	0.57	0.74	1.05	1.06

Table 5AA presents the Rasch difficulty measure before and after anchoring the rating scale threshold parameters of the 1-2, 6-8, and 9-12 grade clusters to the 3-5 grade cluster for Reading. The second column of each table shows the "Original difficulty measure" in logits, when they were independently calibrated (see Chapter 4). The third column, "Anchored difficulty measure," shows the same measures when placed on the 3-5 rating scale. Overall, the difference between the original and the anchored difficulty measure are fairly small, indicating that a single rating scale across all grade-level clusters is a reasonable approach to take.

Table 5AA

Reading Original and Anchored Rasch Difficulty Measures

Cluster	Task name	Original difficulty measure*	Anchored difficulty measure
1-2	R1_A1	-3.18	-3.32
	R2_A2	-0.62	-0.64
	R3_A2	-0.70	-0.73
	R4_A3	-0.43	-0.44
	R5_A3	0.70	0.73
	R6_P1	0.79	0.83
	R7_P1	0.97	1.01
	R8_P2	1.57	1.62
	R9_P2	0.91	0.94
6-8	R1_A1	-3.78	-3.40
	R2_A2	-0.79	-0.76
	R3_A2	-0.98	-0.93
	R4_A3	0.10	0.07
	R5_A3	-0.75	-0.72
	R6_P1	1.27	1.17
	R7_P1	1.19	1.09
	R8_P2	1.79	1.66
	R9_P2	1.95	1.81
9-12	R1_A1	-2.91	-2.87
	R2_A2	-0.78	-0.81
	R3_A2	-0.91	-0.94
	R4_A3	0.03	0.01
	R5_A3	-0.90	-0.93
	R6_P1	1.27	1.29
	R7_P1	0.94	0.95
	R8_P2	1.54	1.56
	R9_P2	1.71	1.74

*See Chapter 5.3

5.4.1.4 Speaking

Table 5BB shows the rating scale statistics for Speaking. Overall, the observed count and observed percent showed that score points ‘0’ and ‘1’ were observed much less frequently than score point ‘2’ and that there is a dramatic increase in frequency from score point ‘1’ to ‘2’. This pattern, although not optimal for obtaining stable rating scale parameters, does reflect the nature of the test administration for Speaking as students had up to six chances to respond to any task.

The observed average measure increased with category value. The observed average measure and expected measure are very close to each other for all score points, indicating a good rating scale model. There was no disordering of step parameters.

There were high (i.e., > 2.00) outfit mean-square statistics for score point '0' on Speaking 3-5 and 9-12. Similar to the findings in Listening and Reading, the high outfit mean-square statistics for these score points were found to be the result of isolated cases of unexpected responses at score point '0' of Task 1. As discussed earlier, Task 1 of the Speaking tests is a very easy task and most students are expected to attain a high score on this task. Therefore, unexpected responses in this task are very likely to produce high outfit mean-square statistics.

Overall, the rating scale statistics showed that the rating scale provided clearly defined and ordered scores and that it is functioning reasonably well. The raw score points also support the construction of a Rasch measure.

Table 5BB
Rating Scale Statistics for Speaking

Cluster	Score point	Obs. count	Obs. percent	Obs. average measure	Expected measure	Original step measure	Final step measure	IN. MSQ	OUT. MSQ
1-2	0	742	24%	-2.43	-2.50	NONE	NONE	1.17	1.34
	1	321	10%	0.51	0.62	-1.11	-1.10	0.92	0.90
	2	2017	65%	4.45	4.43	1.11	1.10	0.94	0.95
3-5	0	669	14%	-2.24	-2.28	NONE	NONE	1.12	2.59
	1	384	8%	0.71	0.76	-1.10	-1.10	0.83	0.86
	2	3675	78%	4.22	4.22	1.10	1.10	1.12	1.32
6-8	0	516	15%	-1.94	-2.13	NONE	NONE	1.23	1.27
	1	272	8%	0.50	0.71	-1.24	-1.10	0.90	0.78
	2	2620	77%	3.93	3.90	1.24	1.10	1.00	1.07
9-12	0	710	20%	-2.10	-2.21	NONE	NONE	1.18	2.39
	1	347	10%	0.60	0.72	-1.24	-1.10	0.88	0.55
	2	2423	70%	3.93	3.91	1.24	1.10	1.11	1.41

Table 5CC presents the Rasch difficulty measure before and after anchoring the rating scale threshold parameters of the 1-2, 6-8, and 9-12 grade clusters to the 3-5 grade cluster for Speaking. The second column of each table shows the "Original difficulty measure" in logits, when they were independently calibrated (see Chapter 4). The third column, "Anchored difficulty measure," shows the same measures when placed on the 3-5 rating scale. Overall, the difference between the original and the anchored difficulty measure are relatively small.

Table 5CC

Speaking Original and Anchored Rasch Difficulty Measures

Cluster	Task name	Original difficulty measure*	Anchored difficulty measure
1-2	S1_A1	-2.66	-2.64
	S2_A2	-1.59	-1.58
	S3_A3	-0.41	-0.41
	S4_A1	-1.21	-1.20
	S5_A2	0.01	0.00
	S6_A3	-0.23	-0.23
	S7_P1	1.52	1.51
	S8_P2	4.57	4.55
6-8	S1_A1	-1.99	-1.91
	S2_A2	-0.89	-0.86
	S3_A3	-0.54	-0.52
	S4_A1	-1.00	-0.96
	S5_A2	-0.73	-0.71
	S6_A3	-0.49	-0.47
	S7_P1	1.52	1.47
	S8_P2	4.12	3.95
9-12	S1_A1	-2.95	-2.82
	S2_A2	-0.84	-0.81
	S3_A3	-0.45	-0.43
	S4_A1	-0.72	-0.69
	S5_A2	-0.49	-0.47
	S6_A3	-0.17	-0.16
	S7_P1	1.81	1.74
	S8_P2	3.81	3.64

*****See Chapter 5.3

5.4.1.5 Writing

Two separate rating scales were used for Writing, one for each of the two parts of the assessment. The tasks in Writing Parts A and B (Tasks 1-8) were modeled using a rating scale with two categories and the tasks in Writing Part C (Tasks 9-10) were modeled using a rating scale with five categories.

Table 5DD shows the rating scale statistics for Writing Parts A and B, Tasks 1-8. Overall, the observed count and observed percent showed that score points ‘0’ and ‘1’ were observed much less frequently than score point ‘2’ and that there is a dramatic increase in frequency from score point ‘1’ to ‘2’. This pattern, although not optimal for obtaining stable rating scale parameters,

once again does reflect the structure of test tasks, which provide multiple opportunities for the student to respond to each task.

The observed average measure increased with category value. The observed average and expected measures are very close to each other for all score points, indicating a good rating scale model. Additionally, there was no disordering in step parameters.

There were high (i.e., > 2.00) outfit mean-square statistics for score point '0' on Writing Parts A and B in grade 6-8 and for score point '1' on Writing Parts A and B in grades 1-2, 3-5, and 9-12. Like other domains, the detailed score point frequencies and fit statistics by task were investigated to identify potential sources of misfit. The investigation showed that unlike other domains, the high outfit mean-square statistics for these score points cannot be isolated to unexpected responses at a particular score point on Task 1. Instead, the sources of misinformation are dependent upon grade-level cluster. For Writing Parts A and B in grades 1-2, 3-5, and 9-12, the misinformation was related to score point '1' of several tasks, including Task 1. For Writing Parts A and B in grade 6-8, the misinformation was related to both score points '0' and '1' of several tasks, including Task 1.

Although the rating scale functions reasonably well when examining the observed average measure and the step parameters, there is some misfit in selected score points of the Writing Parts A and B rating scale for certain grade-level clusters that did not have simple explanations. These high outfit mean-square statistics were not confined to a few explainable responses, suggesting that there may be some issues associated with the definition of the Writing Parts A and B rating scale. Consequently, CAL content developers were asked to review the scoring rules for Writing.

Writing tasks were scored by test administrators using a Writing rubric and an 'Expect' box, which is found in the Student Response Booklet and describes responses that meet task expectations (e.g., trace, copy, write a word). There was variability across test administrators in how they interpreted the rubric and the 'Expect' box. For example, Task 2 across grade clusters asked students to trace a letter. Some test administrators evaluated tracing stringently (expecting the student's tracing to be precisely over the letters being traced) and scored imprecise approximations of tracing as 'Approaches' (i.e., Score point '1' on the rating scale), which identifies students who have not met the task expectations. Other test administrators interpreted a close approximation of the letters being traced as having met the task expectations. Such inconsistency in the interpretation of the Writing rubric and the 'Expect' box could affect the stability of the rating scale. The CAL Test Development team is developing a Writing Scoring Guidance document that aims to address the issue of variability in test administrators' scoring of Writing tasks.

Table 5DD

Rating Scale Statistics for Writing Parts A and B (Tasks 1-8)

Cluster	Score point	Obs. count	Obs. percent	Obs. average measure	Expected measure	Original step measure	Final step measure	IN. MSQ	OUT. MSQ
1-2	0	602	21%	-4.09	-4.34	NONE	NONE	1.17	1.17
	1	622	21%	0.29	0.49	-1.11	-1.10	1.06	2.88
	2	1712	58%	6.91	6.90	1.11	1.10	0.86	0.86
3-5	0	585	13%	-3.16	-3.49	NONE	NONE	1.34	1.60
	1	617	13%	0.57	0.77	-1.10	-1.10	0.98	9.90
	2	3390	74%	7.93	7.92	1.10	1.10	0.88	0.93
6-8	0	340	10%	-2.99	-3.23	NONE	NONE	1.35	9.90
	1	309	9%	0.69	1.03	-1.24	-1.10	0.86	1.36
	2	2607	80%	7.71	7.69	1.24	1.10	0.84	0.91
9-12	0	571	17%	-3.41	-3.49	NONE	NONE	1.03	1.28
	1	401	12%	0.75	0.78	-1.24	-1.10	1.01	9.90
	2	2436	71%	7.38	7.39	1.24	1.10	0.81	0.87

Table 5EE shows the rating scale statistics for Writing Part C, Tasks 9 and 10. Overall, the observed count and observed percent showed that score point ‘0’was observed more frequently than the other score points and that there is a sizeable decrease in frequency from score point ‘0’ to ‘1’. While this pattern seems contrary to those observed from Writing Parts A and B, it reflects the test administration rules. Students would only to move on to Part C of the Writing test if they had scored ‘Meets’ on 7 or 8 tasks in Parts A and B. Hence, many students would not have been administered Part C (based on their performance in Parts A and B) and Test Administrators would have marked ‘Not Administered’ or score point ‘0’.

The observed average measure increased with category value. The observed average measure and expected measure are very close to each other for all score points, indicating a good rating scale model.

There was disordering of step parameters for score point ‘3’ which indicates that this score point was relatively rare. As shown in the observed percent, this score point had proportionally fewer responses than the other score points.

There was a high (i.e., > 2.00) outfit mean-square statistic for score point ‘3’ on Writing Part C in grade 6-8. Like other domains, the detailed score point frequencies and fit statistics by task were investigated to identify potential sources of misfit. It was concluded that these high outfit mean-square statistics were due to some unexpected responses for score point ‘3’ on both tasks of Part C of Writing in grade 6-8. It should be noted that Writing Part C consists of only two

tasks; therefore, the parameter estimates are not expected to be as stable due to the limited number of tasks associated with the rating scale. Overall, the Writing Part C rating scale functions reasonably well.

Table 5EE

Rating Scale Statistics for Writing Part C (Tasks 9-10)

Cluster	Score point	Obs. count	Obs. percent	Obs. average measure	Expected measure	Original step measure	Final step measure	IN. MSQ	OUT. MSQ
1-2	0	474	65%	-2.27	-2.11	NONE	NONE	0.50	0.63
	1	117	16%	-0.31	-0.64	-2.91	-3.24	0.58	0.28
	2	87	12%	0.49	0.41	-0.69	-1.02	0.74	0.58
	3	30	4%	1.12	1.19	1.62	2.22	0.85	0.85
	4	26	4%	2.71	2.71	1.98	2.04	1.53	1.70
3-5	0	483	42%	-6.70	-6.61	NONE	NONE	0.73	0.79
	1	170	15%	-1.99	-2.08	-3.24	-3.24	0.73	0.68
	2	263	23%	0.58	0.51	-1.02	-1.02	0.80	0.81
	3	85	7%	1.55	1.55	2.22	2.22	0.07	1.03
	4	147	13%	2.11	2.06	2.04	2.04	0.93	0.97
6-8	0	239	29%	-6.58	-6.57	NONE	NONE	0.93	0.99
	1	126	15%	-1.92	-1.85	-3.25	-3.24	0.69	0.42
	2	229	28%	0.52	0.52	-1.37	-1.02	0.72	1.47
	3	89	11%	1.69	1.25	1.99	2.22	0.97	3.18
	4	131	16%	2.06	2.59	2.63	2.04	1.29	1.48
9-12	0	340	40%	-6.82	-6.78	NONE	NONE	0.93	1.07
	1	120	14%	-2.32	-2.19	-3.20	-3.24	0.67	0.72
	2	221	26%	0.51	0.48	-1.56	-1.02	0.74	0.65
	3	80	9%	1.67	1.21	2.01	2.22	0.82	0.75
	4	91	11%	2.17	2.61	2.75	2.04	0.97	1.03

Table 5FF presents the Rasch difficulty measure before and after anchoring the rating scale threshold parameters of the 1-2, 6-8, and 9-12 grade clusters to the 3-5 grade cluster for Writing. The second column of each table shows the “Original difficulty measure” in logits, when they were independently calibrated (see Chapter 4). The third column, “Anchored difficulty measure,” shows the same measures when placed on the 3-5 rating scale.

Table 5FF

Writing Original and Anchored Rasch Difficulty Measures

Cluster	Task name	Original difficulty measure	Anchored difficulty measure
1-2	W1_A1	-7.54	-7.05
	W2_A2	-5.05	-4.78
	W3_A3	-1.31	-1.31
	W4_P1	1.98	1.73
	W5_A1	-2.69	-2.63
	W6_A2	-2.21	-2.18
	W7_A3	-0.28	-0.34
	W8_P1	2.85	2.54
	W9_P3	6.98	6.84
	W10_P3	7.28	7.17
6-8	W1_A1	-4.83	-5.95
	W2_A2	-3.43	-4.03
	W3_A3	-1.23	-1.32
	W4_P1	1.21	1.61
	W5_A1	-1.87	-2.08
	W6_A2	-1.96	-2.18
	W7_A3	-0.53	-0.51
	W8_P1	1.40	1.83
	W9_P3	5.78	6.46
	W10_P3	5.46	6.16
9-12	W1_A1	-4.76	-5.6
	W2_A2	-3.55	-4.01
	W3_A3	-1.77	-1.92
	W4_P1	1.36	1.73
	W5_A1	-1.84	-2.00
	W6_A2	-1.64	-1.77
	W7_A3	-0.77	-0.80
	W8_P1	1.44	1.83
	W9_P3	5.78	6.28
	W10_P3	5.76	6.27

*****See Chapter 5.3

5.4.2 Transforming the Logit Scale to the Reporting Scale

Transforming a logit scale to a reporting scale is a simple linear transformation. The procedures to determine the spacing and additive factors for the Alternate ACCESS reporting scale are similar to those described in detail in Kenyon (2006). Each logit score was multiplied by a spacing factor and an additive factor was added. In other words:

New Scale Score = Logit Value * Spacing Factor + Additive Factor.

The equations for converting the logit scale to the ACCESS reporting scale score for the four domains were as follows:

Listening	=	Logit value	*	7.913	+	925.056
Reading	=	Logit value	*	6.026	+	925.788
Writing	=	Logit value	*	2.400	+	926.408
Speaking	=	Logit value	*	4.433	+	924.531

Once the difficulty of the Alternate ACCESS tasks was determined on the reporting scale, a conversion from raw score to scale scores was determined for each student who participated in the Alternate ACCESS operational field test.

5.4.3 Creating the Composite Scores

The scores on the four domain reporting scales were then combined together, weighted as in ACCESS, to create the following four composite scores:

- Oral composite (50% Listening + 50% Speaking)
- Literacy composite (50% Reading + 50% Writing)
- Comprehension composite (30% Listening + 70% Reading)
- Overall composite (15% Listening + 15% Speaking + 35% Reading + 35% Writing).

5.5 Cut Scores

Table 5GG shows the cuts for the four domain scores and four composite scores. The derivation of those cuts is described in the *Alternate ACCESS for ELLs™ Standard Setting Study: Technical Brief* (CAL, 2012a). The column marked "A1/A2" is the cut score between Alternate ACCESS PLs A1 and A2; "A2/A3" the cut score between Levels A2 and A3; and so on. Looking across the first four rows in the table below, within any domain, the cut score increases for higher level proficiency levels.

Table 5GG
Cut Scores

Domain	A1/A2	A2/A3	A3/P1	P1/P2
Listening	925	932	937	942
Reading	924	932	937	942
Speaking	925	930	939	945
Writing	923	931	938	947
Oral Composite	925	931	938	944
Literacy Composite	924	932	938	945
Comprehension Composite	924	932	937	942
Overall Composite	924	931	938	944

6 Standard Setting

The goal of the Standard Setting Study was to interpret performances on the Alternate ACCESS operational field test form in terms of the WIDA ELD Standards, AMPIs, and the WIDA Alternate ELP levels. The study was held in Arlington, VA, on October 9-10, 2012.

6.1 Methodology

The *Angoff Yes/No* methodology was used for all four domains because this method is thought to simplify the cognitive tasks that panelists are asked to perform (Cizek & Bunch, 2007). Having a straightforward cognitive task was important in this study as panelists had to examine many tasks to set four cut scores (A1/A2, A2/A3, A3/P1, and P1/P2) across the four domains (Listening, Speaking, Reading, and Writing).

The *Angoff Yes/No* method was designed for multiple choice and dichotomously scored tasks. This method asks the panelists to consider a student currently functioning at the borderline between two adjacent levels and then to review each question on the test, judging each task as either: a) ***Yes**, the borderline student is **more likely than not to meet expectations for this task***; or b) ***No**, the borderline student is **not** more likely than not to meet expectations for this task.* Under this method, the average of the panelists' *Yes* decisions represents an estimated proportion of the target borderline group who would correctly answer the task.

Some modifications were made to the typical *Angoff Yes/No* methodology. First, for the two tasks in Writing Part C, which are scored using a rubric, panelists were shown various writing samples from all score points and asked to make the decision whether ***Yes**, the borderline student is more likely than not to have produced this sample,* or ***No**, the borderline student is **not** more likely than not to have produced this sample.* This approach to addressing the two rubric-scored tasks meant that the same judging procedures that the panelists used on all other tasks could also be used for these two tasks. The second modification was that the *Yes/No* judgment data collected from the panelists was analyzed using a logistic regression procedure to determine cuts. This approach was used to avoid limitations in the traditional summation approach of calculating

final cut scores with the *Angoff Yes/No* method, which systematically makes lower cuts easier and higher cuts more difficult as compared to the typical Angoff method.

Standards were set on Writing Parts A and B and Speaking using the following procedure. Starting with a student at the lowest borderline as defined and described by the WIDA Alternate ELP levels (i.e., between A1 and A2), panelists independently indicated whether that borderline student would be more likely than not to meet the expectation for the task. If their decision was *No*, panelists then went on to consider a borderline student at the next higher borderline on that same task (i.e., between A2 and A3). This process was continued, considering students at progressively higher levels of proficiency until they reached the highest borderline OR until they indicated *Yes*, that the borderline student would be more likely than not able to meet expectations for that task. Once a decision of *Yes* was made, then all higher borderlines would also necessarily be *Yes* and did not need to be individually considered. This aspect of the procedure greatly simplified the panelists' task.

After panelists considered the borderlines for one task, they then examined the next task and began again by considering the student at the lowest borderline. This process continued until panelists had considered all the borderlines on all the tasks. The test tasks were presented in the same order for consideration as presented in the Alternate ACCESS test booklets. Each panelist completed these evaluations independently. After the first round of evaluations, results for each task were tallied, allowing the panelists to see the 'average' borderline student (e.g., A2/A3) at which the group had determined the task to be more likely than not be answered correctly.

Writing Part C consisted of two writing tasks that were scored using a five-point rubric (i.e., 'No Response,' 'Approaches,' 'Meets 1,' 'Meets 2,' and 'Meets 3') and therefore required a slightly different approach. Sample student responses to the two writing tasks were presented to panelists. Panelists were asked to determine whether a student at each borderline would be more likely than not able to have produced each writing sample.

For Listening and Reading, the prompts for the assessment tasks are repeated to students with increasing levels of support, allowing students multiple opportunities to respond. The repeated prompts are labeled as: CUE A: Initial Prompt; CUE B: Simplified Prompt; CUE C: Simplified Prompt & Answer. A response meeting expectations at CUE A (i.e., with minimal support) is interpreted as demonstrating more ELP than a response meeting expectations at CUE B, while a response meeting expectations at CUE B is exhibiting more proficiency than one at CUE C. For Listening and Reading, the panelists' task was the same as for Writing Parts A and B and Speaking, except that before moving on to the next task they first considered all borderlines on the first task at CUE A, then all borderlines on that task at CUE B, and, finally, all borderlines on that task at CUE C.

For all tasks across all four domains, panelists provided *Yes/No* decisions in a two-round process. In Round 1, panelists independently made their decisions. Staff members then typed the

decisions into a specially prepared Excel spreadsheet which tallied the results by the total number of *Yes* and *No* responses and also presented the percentage of responses in each category. The tallied *Yes/No* decisions across panelists in the group were then revealed to all panelists on a screen with an LCD projector, at which point the panelists had the opportunity to comment on the tallies. Following this discussion, empirical data on student performances on the tasks were presented to the panelists. Using the results from the first round and this new information, the panelists then made a second round of independent *Yes/No* decisions. The Round 2 decisions were again entered and shared with the entire group. A brief opportunity was given to anyone who wanted to comment on the group results before moving on to the next language domain. At the conclusion of the study, researchers used the percentage of *Yes* decisions across panelists from Round 2 to derive the cut scores.

6.2 Panelists

The panels consisted of 34 teachers or administrators from across 20 WIDA Consortium states. These experts are K-12 educators with experience with special education and English Learners. The panelists were recommended for participation by WIDA states, and their qualifications were reviewed by the WIDA Central office at the Wisconsin Center for Educational Research (WCER) at the University of Wisconsin. The name and affiliation of each panelist, ordered alphabetically by state, appears in Table 6A.

Table 6A

Standard-Setting Study Panelists' Names and Affiliations (in Order by State)

Name	State	School or district affiliation
Pamela Bearden	AL	Chilton Country School System
Mayte Cotton	AL	Montgomery Public Schools
Holly Porter	CO	Cherry Creek School District
Siria Rector	CO	St.Vrain Valley School District
Lisa Nelson	GA	Bells Ferry Elementary School, Cobb County School District
Kathy O'Hara Rosa	GA	Clarksdale Elementary, Cobb County School District
Luz Salazar	IL	Booth Tarkington Elementary, Community Consolidated School District 21
Jenny Beltran	IL	Campanelli School, School District 54
Noreen Segal	IL	Community Consolidated School District 21District 21
Beverly Stevens	KY	Clay County Public Schools
Sandy Byrd	KY	Shelby County Public Schools
Kristen Kreit	MD	Shiloh Middle School, Carroll County Public Schools
Sherri Taylor	ME	Maine School Administrative District 15
Alice Habel	MN	Minneapolis Public Schools
Monique El Hani	MN	Minneapolis Public Schools
Yvonne M Field	MT	Montana Office of Public Instruction
Anna Gallo Knight	NH	Hills Garrison School, Hudson School District
Ann Gordon	NH	Oyster River Coop School District
Paddie Donohue	NJ	Lincoln Ave School, Orange Public Schools
Julia T.M. Bowie	NJ	Orange Prep Academy, Orange School District
Robert Romero	NM	New Mexico Public Education Department
Brenda Daw	NV	Clark County School District
Dawn Adams	NV	Washoe County School District
Diane Keene	OK	Lawton Public Schools
Jennifer Bass	OK	Special Services Lawton Public Schools
Adele Claire Wallace	PA	DHH Lengel Middle School, Pottsville Area School District
Jamie Rizzardi	PA	Schuylkill County Schools
Margaret Ames	SD	Madison Elementary, Madison Central School District 39
Julianne Linza	VA	Kilmer Middle School, Fairfax County Public Schools
Donna Hankins	VA	Prince William County Schools
Johanna Snedeker	VT	St. Johnsbury School District
Mercedes Martin	WI	Kromrey Middle School, Middleton Cross Plains Area School District
Valera Crofoot	WY	Carbon County School District
Kimberly Wyman	WY	Washakie County Schools

Demographic information about the panelists was collected and is presented in Table 6B through Table 6E. Table 6B presents the composition of the panelists by gender. All but one of the panelists (97.1%) was female.

Table 6B

Panelists by Gender

	Male	Female	Total
Count	1	33	34
Percent	2.9%	97.1%	100.0%

Table 6C presents the composition of the panelists by race and ethnicity. Although panelists represented a variety of ethnic groups, the vast majority (88.2%) was White and 85.3% were non-Hispanic.

Table 6C

Panelists by Race and Ethnicity

	Ethnicity			Race				
	Hispanic	Non-Hispanic	Total	American Indian	Asian	Black	White	Total
Count	5	29	34	1	1	2	30	34
Percent	14.7%	85.3%	100.0%	2.9%	2.9%	5.9%	88.2%	100.0%

Table 6D presents information on the panel members by years of teaching experience. Panelists were generally very experienced, with a median of experience level of 11-15 years and more than one third of the panelists having at least 16 years of experience.

Table 6D

Panelists' Years of Teaching Experiences

	0-5 years	6-10 years	11-15 years	16-20 years	21+ years	Missing	Total
Count	5	7	9	8	4	1	34
Percent	14.7%	20.6%	26.5%	23.5%	11.8%	2.9%	100.00%

Note: One panelist did not provide this information

Table 6E presents information on the highest level of education attained by the panel members. As shown, 33 panelists (97.1%) had completed at least some graduate study. Of these, 25 panelists (73.5%) held a Master's degree and four panelists (11.8%) had completed some doctoral study or held a doctoral degree.

Table 6E
Panelists' Highest Level of Education

	Highest Level of Education					
	Bachelor's degree	Some graduate study	Master's degree	Some doctoral study	Doctoral degree	Total
Count	1	4	25	1	3	34
Percent	2.9%	11.8%	73.5%	2.9%	8.8%	100%

In summary, the panelists represented the variety of states in the WIDA Consortium well, had considerable teaching experience, and were well-educated. They appeared well-qualified to serve as panelists in the study.

6.3 General Procedures

6.3.1 Panel Assignments

The study took place on October 9-10, 2012 in Arlington, VA. Panelists were pre-assigned to serve on one of the four grade-level cluster panels (1-2, 3-5, 6-8, 9-12) as well as one of the four groups that created borderline descriptors by domain (Listening, Reading, Speaking, Writing). Panelists were assigned to a grade-level cluster that they were either currently teaching or had taught in the past. In addition, assignments were made with the aim of having a variety of states represented in each group. There were eight or nine panelists at each grade-level cluster panel and within each domain group.

6.3.2 Materials

Each panel member received reference materials, including a copy of the Alternate ACCESS Performance Level Descriptors, grade-level cluster AMPIs, and Grade Cluster Test Materials. After developing Borderline Descriptors, this information was also distributed. For Writing Part C, in addition to the Grade Cluster Test Materials, six written student responses to each of the two rubric-scored tasks were presented to the panelists for a total of 12 writing samples. Student writing samples were selected based on their teacher-assigned raw score points. An attempt was made to select at least one writing sample at each of the four raw score point levels ('Approaches,' 'Meets 1,' 'Meets 2,' and 'Meets 3') such that all four raw score points were presented. Since the same writing tasks are presented in grade clusters 1-2 and 3-5 and in 6-8 and 9-12, only two sets of writing samples were used: one for the 1-2 and 3-5 clusters and another for the 6-8 and 9-12 clusters.

In addition to the reference materials, panelists received a Panelist Standard Setting Booklet, which included a demographic questionnaire, practice worksheets, domain specific worksheets, and evaluation questionnaires for each domain. The demographic questionnaire was completed by panelists as they arrived. Panelists completed the evaluations after each training session and

after completing each domain to provide feedback on the training, process, and materials both generally and for each domain. All panelists' work was completed and recorded in these booklets.

6.3.3 Day 1 Procedures

After introductions, Dr. Dorry Kenyon of CAL and Erin Arango-Escalante of WIDA led the panelists through group training, providing them with background on the student population. The training began with an explanation of the goals of the study and a walk-through of the general logic of standard setting. Then, each component of the Standard Setting Study was addressed in turn. Focusing on borderline students, panelists worked within their preassigned language domain working groups (Listening, Reading, Speaking, or Writing) to develop Borderline Descriptors for each division between proficiency levels for that domain. The Borderline Descriptors were then shared with the entire group of panelists. After the Borderline Descriptors were agreed upon for each domain, Deepak Ebenezer of CAL trained the panelists' on the test content and administration procedures of Alternate ACCESS. Dr. Kenyon then led the panelists through the process of making *Yes/No* decisions and indicating their decisions in the Panelist Standard Setting Booklet. Panelists practiced making several decisions and had opportunities to ask questions about the assignment. After the training, panelists moved into separate rooms according to their assigned grade-level cluster panel and worked on Writing until the end of Day 1. Because Writing Part C required a slightly different procedure than Writing Parts A and B, Writing standard setting activities were separated into two parts, I and II. In Writing Part I, panelists developed ratings for Writing Parts A and B; in Writing Part II, panelists developed Writing Part C ratings. Two staff members from CAL or WIDA worked with each panel, one as the group facilitator and one as an assistant. Dr. Kenyon moved from room to room to monitor the work and to answer questions. Answers to all questions were shared with all rooms.

6.3.4 Day 2 Procedures

On the second day, the panelists first met as a group to be trained on the Listening and Reading processes. During this session, the panelists practiced the *Yes/No* decision process, considering responses at CUE A, CUE B, and CUE C for each task. Because of this cueing structure, the standard setting process and data recording sheets for Listening and Reading were slightly more complicated than for other domains and, therefore, required additional training in order to familiarize the panelists with the procedures. After the training, the panelists split into their grade-level cluster assignments and worked on the domains of Speaking, Listening, and Reading, following the same procedures as they did for Writing on Day 1.

6.4 Analyses and Results: Panelists' Evaluations

Since standard setting is a socially-moderated event, it is important to consider participants' evaluation of the process, as well as the outcomes. Part of the validity evidence for the standard setting process is ensuring that the panelists felt that they were well-trained and prepared for the Standard Setting. Thus, at the end of each training session, panelists completed an evaluation

form which provided them the opportunity to evaluate all aspects of the Standard Setting training.

The panelists evaluated several aspects of the training by responding to Likert scale items using response levels of 1 to 4, with 4 being the highest rating. The items were the following:

1) The logic and the goals of the Standard Setting were clear.
2) The student population and participation criteria were clearly explained.
3) The overview of the WIDA Alternate ELD Standards was clear and helpful.
4) The overview of the Alternate ACCESS test was clear and helpful.
5) The lead facilitator clearly explained the task.
6) The training and practice helped me understand how to perform the task.
7) The facilities and the food service helped create a productive working environment.

For each aspect, Table 6F shows the average rating, standard deviation, and the number of ratings. All aspects of the training received high ratings, with panelists assigning the highest ratings to the explanation of the student population and participation criteria (M = 3.53) and the lead facilitator's explanation of the task (M = 3.53). The training and practice (M = 3.47) and the overview of the Assessment (M = 3.44) received the next highest set of ratings. Ratings increased after the Day 2 refresher training and practice; both the lead facilitator (M = 3.66) and the training and practice (M = 3.59) received higher ratings on day 2.

Table 6F
Panelists' Evaluation of Training

Evaluation	Statistic	Logic and goals of standard setting	Student population and participation criteria	Overview of WIDA Alternate ELD Standards	Overview of Alternate ACCESS Assessment	Lead facilitator	Training and practice	Facilities and food
Day 1 Evaluation	Mean	3.35	3.53	3.35	3.44	3.53	3.47	3.73
	Std Dev	0.69	0.51	0.60	0.50	0.51	0.56	0.45
	N	34	34	34	34	34	34	33
Day 2 Refresher Evaluation	Mean					3.66	3.59	
	Std Dev					0.48	0.50	
	N					32	32	

After completing the standard setting process for each domain (Listening, Reading, Speaking, Writing Part I, and Writing Part II), the panelists evaluated their experience. Using the same 4-point Likert scale, the panelists provided responses to the following domain-specific items:

1) The Performance Level Descriptors, AMPIs, and Borderline Descriptors helped me perform the task.
2) The materials used to make *Yes/No* decisions were clear and helpful.

3) The process used to make *Yes/No* decisions was clear.
4) There was adequate time to perform my task.
5) The group facilitator kept me on task.
6) The group discussion after Round 1 was helpful.
7) I am confident that the Round 2 data adequately reflects the performance of borderline students.

For each domain, Table 6G shows the average rating, standard deviation, and the number of ratings across the four grade-level panels.

On the scale of 1 to 4, all aspects of the domain work received high ratings. The results indicate that the panelists felt that they had adequate time to complete the tasks (range: 3.70 - 3.79), that the Round 2 group discussions were helpful (range: 3.47 - 3.73), and that the group facilitator kept the group focused on the task (range: 3.65 - 3.73). The panelists had high confidence that the Round 2 data adequately reflected the performance of borderline students (range: 3.41 – 3.65). Not surprisingly, the panelists were most confident about the results of Speaking, the last domain they completed (M = 3.65), and were least confident about the results of Writing Part I, the first domain they completed (M = 3.41). However, all ratings were high across the board, providing support for the validity of the standard setting process.

Table 6G
Panelists' Evaluation by Domain

Domain	Statistic	Descriptors & AMPIs	Working materials	The process	Adequate time	Group facilitator	Round 1 Group Discussion	Confidence in Round 2 results
Listening Evaluation	Mean	3.56	3.53	3.53	3.71	3.65	3.71	3.52
	Std Dev	0.70	0.51	0.51	0.46	0.49	0.46	0.51
	N	34	34	34	34	34	34	31
Reading Evaluation	Mean	3.58	3.55	3.64	3.76	3.73	3.70	3.55
	Std Dev	0.61	0.51	0.49	0.44	0.45	0.47	0.51
	N	33	33	33	33	33	33	33
Speaking Evaluation	Mean	3.74	3.74	3.71	3.79	3.71	3.71	3.65
	Std Dev	0.51	0.45	0.46	0.41	0.46	0.46	0.49
	N	34	34	34	34	34	34	34
Writing Evaluation Part 1	Mean	3.59	3.50	3.53	3.76	3.68	3.47	3.41
	Std Dev	0.56	0.51	0.56	0.43	0.53	0.51	0.50
	N	34	34	34	34	34	34	34
Writing Evaluation Part II	Mean	3.52	3.48	3.52	3.70	3.67	3.73	3.45
	Std Dev	0.62	0.51	0.51	0.47	0.54	0.45	0.56
	N	33	33	33	33	33	33	33

6.5 Analyses and Results: The Cut Scores

On ACCESS (Kenyon, 2006), which is used by the general ELL student population, cut scores are provided by grade level. Because the cognitive abilities of general ELL students are expected to increase with age, it is reasonable to require students from higher grades to qualitatively demonstrate more language than students in lower grade levels, even though their general level of ELP may be considered the same.

For Alternate ACCESS, however, it was determined that its student population is so unique that a different approach was needed. In students with severe cognitive disabilities, the cognitive abilities that support language proficiency development are not expected to increase dramatically from one grade level to the next. At this point in the understanding of the development of ELP in such students, it appears more appropriate to use the same cut scores for all grade clusters (from grades 1 to 12) by domain. In this way, it will be easier to detect growth in ELP from year to year for this population of English learners.

To derive common cut scores by domain, a series of logistic regression analyses were conducted. Logistic regression is a statistical technique for relating a continuous variable (i.e., the difficulty of the assessment tasks) to a dichotomous outcome (i.e., the *Yes/No* decisions made by the panelists). A logistic regression analysis was conducted for each cut for each domain (e.g., the A3/P1 cut for Listening) using the panelists' *Yes/No* decisions across test tasks and grade clusters in that domain. The logistic function was used to find the location along the underlying ability continuum at which 50% of the panelists thought that the borderline student is more likely than not to meet the task expectations. This point became the cut point between the two adjacent proficiency levels being analyzed.

Thus, when conducting the analysis between levels, there were two sets of variables for each task: a) the *location parameter* of the task's difficulty, which is used as the indicator of task location on the student ability continuum; and b) the percentage of judges indicating ***Yes***, *the borderline student is more likely than not to meet task expectations*. The location parameter for each task corresponds to the Rasch-Thurstone thresholds on the rating scale, which dichotomize the rating scale at each category boundary into a 50% probability of being observed below the category and a 50% probability of being observed in or above the category (or score point) for the task. This value is a function of both the task difficulty parameter and a threshold parameter, and it is the ability level associated with a 50% chance of obtaining the particular score point or better on the task.

Because all test tasks across grade clusters by domain needed to be analyzed together in one logistic regression analysis, the location parameters needed to be on a common scale. Task parameters across grade clusters by domain were put on a common Rasch logit scale using the rating scale parameters for the 3-5 grade cluster as anchors. (Refer to *Alternate ACCESS for ELLs™ Series 100 Development and Operational Field Test: Technical Brief* [CAL, 2012b] for

additional information about scaling.) After the scaling was completed, location parameters were obtained and used in the logistic regression analysis.

Figure 6A provides an example of a logistic regression procedure. This figure shows the results for the borderline cut between A3 and P1 for Listening. The vertical axis represents the probability of the panelists saying ***Yes***, *a student at the A3/P1 borderline is more likely than not to meet task expectations*. The horizontal axis represents the location parameters on the Rasch logit scale. The 108 dots in Figure 1 represent each observation made by the panelists across grade levels and across the three levels of cueing for nine Listening tasks. Each task is positioned at its location parameter on the difficulty scale and at the percent of panelists indicating ***Yes*** for that borderline student on a task located at that difficulty level. It is clear that for the easiest tasks (the left side of the horizontal axis), 100% of the panelists indicated ***Yes***, *the borderline A3/P1 borderline student is more likely than not to meet task expectations.* As the tasks increase in difficulty, the percentage of panelists indicating *Yes* likewise decreases. The dot to the furthest right indicates the most difficult task, to which 100% of the panelists indicated ***No***, *the A3/P1 borderline student is not more likely than not to meet task expectations.*

To find the point at which at least 50% of the panelists thought that the A3/P1 borderline student is more likely than not to meet the expectations of a Listening tasks, it is necessary to find 0.5 on the vertical axis, follow the horizontal line across to the point where it meets the curve, and go down to find the corresponding location parameter on the horizontal axis. This location parameter represents the cut between level A3 and level P1 for the Listening domain; that is, the point where panelists would say the borderline student has a 50% probability of meeting/not meeting task expectations.

Computationally, after the parameters of the logistic regression function are estimated, finding the cut score is straightforward. In this illustrated example, for the Listening A3/P1 cut score analysis, the location parameter that corresponds to the 50% probability level is 1.58 logits. On a Listening task at this level of difficulty, about 50% of the panelists would indicate *Yes* and 50% of the panelists would indicate *No.* This location parameter can also be interpreted as an ability parameter. From that parameter, the cut score is determined by converting the logit value to the Alternate ACCESS score scale.

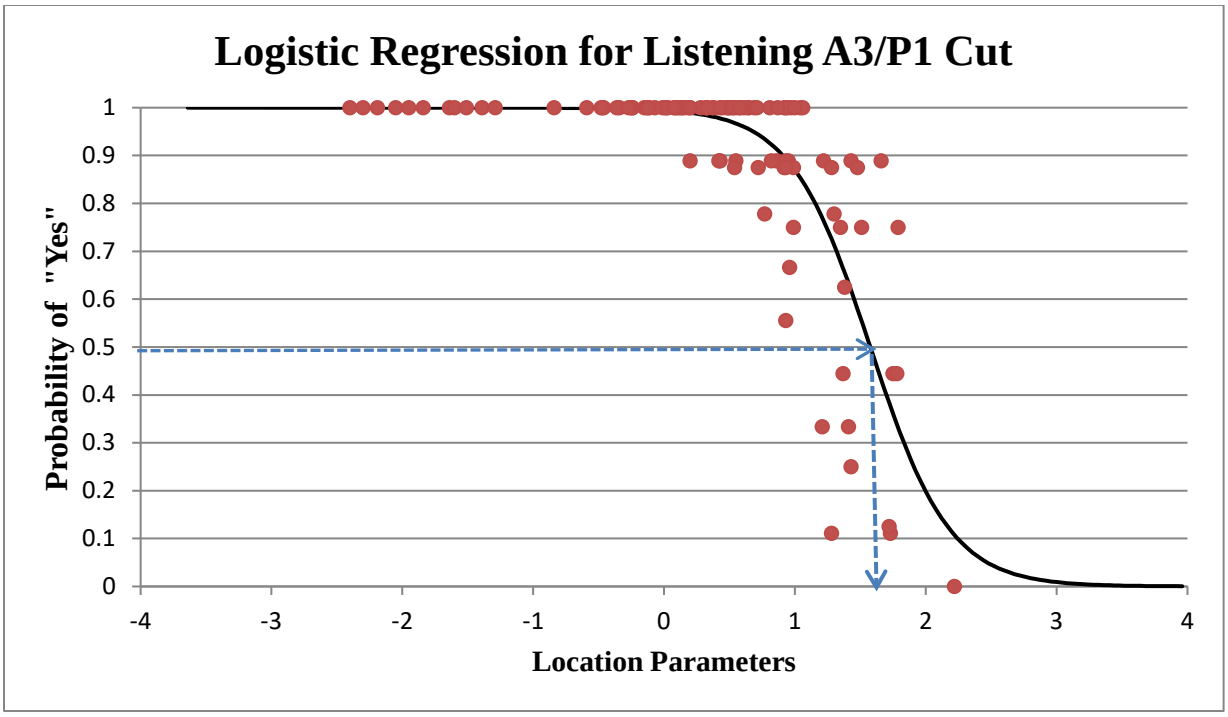

Figure 6A: Illustration of Cut Score Selection Using Logistic Regression

Using the aforementioned logistic regression procedure, cut scores were obtained for the Alternate ACCESS proficiency levels. In addition to the cut scores for each domain, cut scores were determined for the four composite scores (Oral Language, Comprehension, Literacy and Overall) by means of the same weightings used for ACCESS. The Oral Language cut scores were determined by averaging the cut scores for Listening and Speaking, and then truncating to the nearest integer. Similarly, the Literacy cut scores were determined by averaging the cut scores for Reading and Writing, and then truncating to the nearest integer. The Comprehension and Overall cut scores were determined using the following weightings:

Comprehension = .30 * Listening + .70 * Reading
Overall = .15 * Listening + .35 * Reading + .35 * Writing +. 15 * Speaking

The Comprehension and Overall results were then truncated to the nearest integer.

Table 6H presents the cuts for the four domain scores and four composite scores. The column marked “A1/A2” is the cut score between Alternate ACCESS PLs A1 and A2; “A2/A3” is the cut score between PLs A2 and A3; and so on. Looking across the first four rows in the table below, within any domain, the cut score increases at higher level proficiency levels, indicating that students who achieve higher scale scores also receive higher proficiency level scores.

Table 6H

Cut Scores on Alternate ACCESS

Domain	A1/A2	A2/A3	A3/P1	P1/P2
Listening	925	932	937	942
Reading	924	932	937	942
Speaking	925	930	939	945
Writing	923	931	938	947
Oral Composite	925	931	938	944
Literacy Composite	924	932	938	945
Comprehension Composite	924	932	937	942
Overall Composite	924	931	938	944

7 Initial Investigations of Test Validity

7.1 Overview

"Validity refers to the degree to which evidence and theory support the interpretations of test scores by proposed users of tests. Validity, therefore, is the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999, p. 9). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property; rather, it is an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment.

In the past two decades, argument-based approaches (Kane, 1992, 2006) to validation have emerged. The Assessment Usage Argument (AUA) (Bachman, 2005) provides a convenient framework for guiding test development and assessment usage justifications. Central to the AUA approach to validity is to establish a conceptual framework consisting of a series of inferences that link the test taker's performance to a claim along with warrants and backing to support the claim. In the following sections, we describe two claims that were investigated, the specific statements that elaborate and support the claims (***warrants***), the data and evidence collected to support the claims (***backings***), and the methodology used to evaluate the validity of the claims. These two claims address test validity from the perspective of score interpretations and score consistency. They correspond to the bottom half of the data-claim inferential link illustrated by Bachman and Palmer (2010, Figure 5.4) as the basic 'building block' of AUA.

7.2 Claim 1: Interpretations of Scores

Claim 1: Alternate ACCESS proficiency scores support appropriate and meaningful interpretations of students' abilities and ELP levels in terms of the Alternate WIDA ELD Standards.

The justification for making such an inference is referred to as a ***warrant***: The interpretations of students' performance on Alternate ACCESS are meaningful when described in terms of the Alternate ACCESS proficiency levels. These proficiency levels can be properly interpreted by using the Alternate WIDA ELD Standards for students with severe cognitive disabilities.

Backing to support this claim comes from the rigorous test development procedures (Backing 1.1) that ensure proper alignment of the test tasks to the Alternate WIDA ELD Standards and the AMPIs, from psychometric analyses examining the construct assessed by the Alternate ACCESS tasks (Backing 1.2), and from the results of two concurrent validity studies (Backing 1.3 and 1.4).

Backing 1.1: Test development procedures provide evidence of the content validity of Alternate ACCESS. The WIDA ELD Standards (i.e., Social and Instructional Language, Language of Language Arts, Language of Mathematics, Language of Science, and Language of Social Studies)—which are grounded in scientifically based research on best practices in general, English as a Second Language, and bilingual education—guided the development and task specifications for Alternate ACCESS. Every task on Alternate ACCESS was developed to target at least one of the five WIDA ELD Standards. And each task on Alternate ACCESS was developed to target one of the AMPIs.

Additional evidence of content validity is provided by a series of qualitative evaluations of Alternate ACCESS test content during the Alternate ACCESS test development and analysis stage by content experts and stakeholders.

The first was the expert panel review, whose recommendations on the format and content of the test guided task development throughout the process. That panel was reconvened after the initial cognitive labs to review the materials and comment on how engaging the tasks were, whether they were developmentally appropriate, and how accessible the language of the tasks was.

The next review was by the Bias and Sensitivity panel, which was comprised of teachers and experts in special education. The purpose of this review was to ensure that the tasks were appropriate and accessible to the students in the target population and that they aligned with the sensitivity guidelines.

The next review was performed by WIDA and SEA representatives. Following the copy editor's review of materials, in December 2011, a forms review was conducted at CAL with

representatives from WIDA and SEAs, as well as specialists from the special education field. The information from the forms review was used to further refine the test forms.

After the operational field test, the Standard Setting committee, comprising 34 teachers from 20 WIDA Consortium states, met to interpret scores on the test in terms of the WIDA ELD levels. The committee was able to do so with a high degree of agreement and confidence. (An overview of the standard setting results can be found in the *Alternate Access for ELLs™ Standard Setting Study: Technical Brief*; CAL, 2012a).

The knowledge, expertise, and professional judgments of the experts ultimately ensured that the content of ACCESS formed a legitimate basis upon which to validly derive conclusions about students' ELP.

Backing 1.2: Construct validity for Alternate ACCESS is evaluated with the use of Rasch analyses (see Chapter 5.3). Items that fit the Rasch model are likely to be measuring the intended construct of ELP and to contain little construct irrelevance. Construct validity for the Writing section is evaluated with the use of rater reliability analyses (see Chapter 7.3) to indicate that raters used the scoring procedures and training materials to render reliable scores for students' writing samples.

Rasch models are confirmatory and assume a strong theoretical grounding for item analyses. One major threat to construct validity is the inevitable inclusion of construct-irrelevant variance. These are variances that are related to sub-dimensions of abilities measured by the test items and are irrelevant to the focal construct. Thus, measures that fit our measurement model may be considered, psychometrically speaking, very strong measures. Rasch analysis is also a powerful tool for evaluating construct validity. The items that do not fit the Rasch model may indicate instances of multidimensionality. The items that fit are likely to be measuring the single dimension intended by the construct. Therefore, misfitting items may be indications of construct-irrelevant variance. As presented in Chapter 5.4, 142 of the 144 total tasks across all the grade-level clusters and domains of the Alternate ACCESS tasks fit the Rasch model well in terms of infit (which is a more reliable gauge of model fit, as infit statistics are more resistant to influence by outliers) and are productive for measurement according to the statistical criterion used to flag tasks for review. These results are a strong indication that the Alternate ACCESS tasks measure the construct that the tests were designed to measure. In addition, while some of the outfit statistics were inflated as a result of inconsistent or unexpected responses, this could well be expected in the first administration of this assessment, particularly in light of the population that is being tested. It expected that after tightening the writing rubric, more training, and more experience with this assessment, there will be fewer outliers that inflate the outfit statistics, especially on the easiest items or least observed score categories.

For Writing, rater reliability analyses (Chapter 7.3) suggest that the score variability associated with the raters was minimal and that the scoring procedures and training materials were sufficient for the raters to render reliable Writing scores. This supports the claim that the construct-irrelevant variance related to scoring the Writing responses was minimized.

Backing 1.3: Relationships between domain scores provide some evidence of the concurrent validity of Alternate ACCESS. Because Alternate ACCESS is designed to measure ELP in all domains, and because that general proficiency should underlie proficiency in the individual domains, we would expect a moderately high correlation between the scale scores in the individual domains. More specifically, we would expect related domains such as Listening and Speaking or Reading and Writing to show a relatively high correlation.

Table 7A shows the intercorrelations of the domain scale scores for the 1,876 students who participated in the field test. Nonresponses, defined by blank response records, were removed from the correlations by domain, so the total number of students used to calculate the correlations varies by domain (see Table 7B below for the number or students in each domain). Overall, correlations were moderate to high between domains and ranged from .74 between Listening and Writing to .84 between Listening and Reading. As expected, Reading and Writing, which both measure literacy skills, showed a high correlation (r = .80). Listening unexpectedly correlated more highly with Reading (r = .84) than with Speaking (r = .75). Although we might expect Listening to correlate most highly with Speaking (since they both measure oral proficiency skills), this result may be due to a method effect since Listening and Reading had very similar task specifications. For example, both were selected response, both domains had similar scoring rules, and both allowed for repetition of the tasks with increasing levels of support, allowing students multiple opportunities to respond.

Table 7A
Correlations of Scale Scores by Domain

	Listening	Reading	Speaking	Writing
Listening	1			
Reading	.84**	1		
Speaking	.75**	.76**	1	
Writing	.74**	.80**	.78**	1

**p < .01

Backing 1.4: Additional concurrent validity for Alternate ACCESS was obtained using data collected from a special research study conducted concurrently with the Series 100 field test administration. Details of the study can be found in the *Alternate Access for ELLs™ Series 100 Teacher Rating Worksheet: Technical Report* (CAL, 2013). The first part of the study asked teachers to evaluate students' development of English language as defined by the WIDA Alternative ELD standards using a Teacher Rating Worksheet. The Worksheet contains a

checklist of English language descriptors by domain that had been aligned to the Alternate ELP Levels A1-P2. Teachers were instructed to rate student performance on classroom activities in relation to these descriptors by domain. Since these descriptors and the Alternate ACCESS test are both based on the WIDA Alternate ELD standards, we would expect a positive relationship between student performance on the descriptors as rated by the teachers and his or her actual performance on Alternate ACCESS.

Teacher ratings of students' performance on the descriptors were first scaled using a Rasch model. Based on the teacher ratings, each student was assigned a logit score. Each student's scale score on Alternate ACCESS was also computed. These two sets of scores were correlated to examine the extent to which the students' ability in performing the descriptors, as rated by their teachers, correlated with their test performance on Alternate ACCESS.

Overall, correlations were positive and moderate to moderately high, ranging from .68 for Listening to .78 for Writing. The results suggest that teachers, when using the Teacher Rating worksheets, assess their students' developing proficiency in English in a manner that is similar to student performance assessment in Alternate ACCESS. The finding that teacher ratings of their students' developing ELP, based on their daily interactions with students, are similar to the performance evaluations on the Alternate ACCESS assessment, provides some support for the claim that Alternate ACCESS appropriately measures the ELP construct as defined in the WIDA Alternate ELD Standards.

7.3 Claim 2: Consistency of Scores

Score consistency refers to the extent to which test takers' performances on different assessments of the same construct yield the same result (Bachman & Palmer, 2010). A consistent assessment will provide essentially the same information about test takers' abilities across different aspects of assessment conditions, like different test items, different test administrations, different times, or different raters.

Score consistency can be affected by many factors, such as test takers' psychological or physical state, the administering of alternate test forms that contain different items, environmental factors like room conditions, test administrators' differences in administration procedures, and rater judgments of the test takers' responses or performance. WIDA cannot control all these factors in a given test situation but has taken steps to ensure score consistency.

WIDA has strived to reduce the chance of measurement error in the items and test forms by designing tests that contain a large-enough sample of high-quality items in order to better sample students' performance. This ensures that students would receive similar scores on the test over repeated test administrations. However, score consistency is a matter of degree and needs to be examined using empirical data. The degree of score consistency for Alternate ACCESS was examined using measures of test reliability. The *Standards for Educational and Psychological*

Testing (AERA, APA, & NCME, 1999) defines reliability as "the consistency of [educational] measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). Analysis of test reliability provides information about the likelihood that students would receive the same score on the test over repeated test administrations.

Claim 2: Test takers' performances on Alternate ACCESS are consistent across different aspects of assessment conditions.

Warrant: Demonstrated by the accepted use of computed statistics and supported by effective rater training procedures, Alternate ACCESS tests produce scores that are consistent across different test administrations, different test tasks, and (for the Writing tasks) different raters.

Backing to support this claim comes from analyses of Alternate ACCESS test administration (Backing 2.1) and from psychometric and statistical analyses of test reliability and rater reliability (Backing 2.2 and 2.3).

Backing 2.1: Test administration procedures are standardized to reduce the chance of measurement error due to differences in administration procedures.

WIDA has attempted to address environmental factors by specifying to test administrators the room setup, appropriate amounts of light and noise, desk arrangements, duration of testing times, and security of materials, among other things. To minimize differences in administration procedures and in rater variation on the Writing and Speaking sections, WIDA has produced the following training materials for test administrators: a) the *WIDA Alternate ACCESS for ELLs™ Test Administration Manual* (WIDA, 2012a), which details how to prepare for, administer, score, and interpret scores on Alternate ACCESS; b) the *WIDA Alternate ACCESS for ELLs™ Test Administration Training Tutorial* (WIDA, 2012b), which covers general information on the structure of Alternate ACCESS and shows scenes demonstrating the administration; and c) additional sources on the WIDA ACCESS website, including a toolkit and various webinars that are available for people preparing to administer Alternate ACCESS.

Backing 2.2: Empirical analysis of test reliability provides evidence that students would receive similar scores on the test over repeated test administrations (assuming that no additional learning has taken place).

The test reliability information reported in Table 7B is based on Classical Test Theory (Crocker & Algina, 2009). This table shows the number of students, the number of tasks, the Cronbach's alpha, and the classical standard error of measurement (SEM) for each grade-level cluster form. Cronbach's alpha is a numerical coefficient that is widely used as an estimate of test reliability. It expresses how well the tasks on a test appear to measure the same construct (i.e., the internal

consistency of test tasks). Nonresponses were removed from the overall total by domain to calculate these reliability indices, as nonresponses would have artificially inflated Cronbach's alpha and because no information was available about these students on these domains. Nonresponses were identified as three consecutive "Blank" ratings assigned to the first three tasks. All other responses were included in the reliability analyses. The *standard error of measurement* (SEM) presented in Table 7B is based on classical test theory. Unlike IRT, in this approach, SEM is seen as a constant across the spread of test scores (ability continuum). Thus, it is ***not*** conditional on ability being measured. It is, however, a function of two statistics: the reliability of the test and the (observed) standard deviation of the test scores. It is calculated as

$$\text{SEM} = SD\sqrt{1 - reliabilty}$$

+From Table 7B, we see that the Cronbach's alpha value for each domain in each grade-level cluster of Alternate ACCESS is above .90, which is considered excellent (George & Mallery, 2003). For Listening, Reading, and Speaking, Cronbach's alpha was calculated based on all tasks. For Writing, the Cronbach's alpha was calculated based on the first 8 tasks included in Writing Parts A and B, because they were scored using the same scoring rules. Writing Part C, which includes tasks 9 and 10 for Writing, used different scoring rules and therefore was not included in this analysis.

Table 7B
Reliability of Alternate ACCESS Series 100

Grade-level cluster	No. of students	No. of tasks	Cronbach's alpha	SEM
Listening				
1-2	387	9	.95	0.034
3-5	595	9	.95	0.022
6-8	431	9	.95	0.009
9-12	446	9	.95	0.019
Reading				
1-2	388	9	.95	0.060
3-5	593	9	.96	0.024
6-8	433	9	.95	0.045
9-12	442	9	.94	0.059
Speaking				
1-2	385	8	.97	0.010
3-5	591	8	.96	0.006
6-8	426	8	.97	0.004
9-12	435	8	.97	0.006
Writing				
1-2	367	8	.95	0.020
3-5	574	8	.96	0.008
6-8	407	8	.96	0.004
9-12	426	8	.96	0.005

Backing 2.3: Analysis of rater reliability provides evidence that the rating process and the training materials are working as intended and that the agreement levels among raters are adequate.

For the Writing domain, interrater reliability data collected from a special scoring study conducted at CAL were analyzed and are presented in this section. This special study was undertaken because the Writing tasks and accompanying rubrics were newly developed and therefore required close scrutiny in terms of reliability.

While the interrater reliability of the Speaking domain is also an important area of concern, no special study on the interrater reliability of the Speaking subtest has yet been conducted because that part of the assessment is scored in the same manner as ACCESS.

The Writing tasks were scored by administrators during the Alternate ACCESS test administration. Writing Parts A and B, Tasks 1-8, were scored on a two-point scale, and Writing Part C, Tasks 9-10, were scored on a four-point scale. See Table 5-B for more details about possible raw scores for the Writing Tasks.

A special internal study was conducted by two CAL test developers familiar with Alternate ACCESS to explore the reliability of the Writing rubric and to develop the Writing Scoring Guidance document. The two raters independently scored samples for each grade-level cluster for the eight writing tasks in Writing Parts A and B and the two Writing tasks in Writing Part C. The results of the independent rating study for Writing Parts A and B are presented in Table 7C. The table includes the number of sample responses scored by the two raters for each task. The table summarizes the number and percent of responses that showed agreement between the two raters. As can be seen from Table 7C, the two raters agreed on 95% of sample responses.

Table 7C

Reliability of Writing Parts A and B Rubric by Task

Grade- level cluster		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Total	% of Total
1-2	#Responses	1	1	1	2	1	1	1	2	10	100%
	#Agreement	1	1	1	2	1	1	1	2	10	100%
	#Adjacent	0	0	0	0	0	0	0	0	0	0%
3-5	#Responses	1	1	1	2	1	1	1	2	10	100%
	#Agreement	1	1	1	2	1	1	0	2	9	90%
	#Adjacent	0	0	0	0	0	0	1	0	1	10%
6-8	#Responses	1	1	1	2	1	1	1	2	10	100%
	#Agreement	1	1	1	2	1	1	1	2	10	100%
	#Adjacent	0	0	0	0	0	0	0	0	0	0%
9-12	#Responses	1	1	1	2	1	1	1	2	10	100%
	#Agreement	1	1	1	2	1	1	1	1	9	90%
	#Adjacent	0	0	0	0	0	0	0	1	1	10%
Total	#Responses	4	4	4	8	4	4	4	8	40	100%
	#Agreement	4	4	4	8	4	4	3	7	38	95%
	%Agreement	100%	100%	100%	100%	100%	100%	75%	88%	95%	
	#Adjacent	0	0	0	0	0	0	1	1	2	5%
	%Adjacent	0%	0%	0%	0%	0%	0%	25%	13%	5%	

The results of the independent rating study for the two tasks in Writing Part C are presented in Table 7D. These tasks were scored out of a possible four points, so there was more variation in the rater's assigned scores than for Writing Parts A and B. The two raters agreed on 84% of the sample responses; the remaining scores differed by one point. Thus, when exact agreement and adjacent agreement are considered together, the two raters agreed on 100% of the sample responses.

Table 7D

Reliability of Writing Part C Rubric by Task

Grade-level cluster		Task 9	Task 10	Total	% of Total
1-2	# Responses	8	8	16	100%
	# Agreement	6	7	13	81%
	# Adjacent	2	1	3	19%
3-5	# Responses	8	8	16	100%
	# Agreement	7	6	13	81%
	# Adjacent	1	2	3	19%
6-8	# Responses	8	8	16	100%
	# Agreement	7	5	12	75%
	# Adjacent	1	3	4	25%
9-12	# Responses	8	8	16	100%
	# Agreement	8	8	16	100%
	# Adjacent	0	0	0	0%
Total	# Responses	32	32	64	100%
	# Agreement	28	26	54	84%
	% Agreement	88%	81%	84%	
	# Adjacent	4	6	10	16%
	% Adjacent	13%	19%	16%	

8 References

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

Andrich D. (1978) A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F & A. B. Palmer (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. UK: Oxford University Press.

Center for Applied Linguistics. (2012a). *Alternate ACCESS for ELLs™ Standard Setting Study: Technical Brief.* Available at www.wida.us.

Center for Applied Linguistics. (2012b). *Alternate ACCESS for ELLs™ Series 100 Development and Operational Field Test: Technical Brief.* Available at www.wida.us.

Center for Applied Linguistics. (2013). *Alternate ACCESS for ELLs™ Series 100 Teacher Rating Worksheet: Technical Report.* Available at www.wida.us.

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

Crocker, L., & Algina, J (2009). *Introduction to Classical and Modern Test Theory*. Cengage Learning: Mason, Ohio.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (2006). Validation. In R. Brennan (Ed.) *Educational Measurement* (4th ed.) Westport, CT: American Council on Education and Praeger Publishers.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Kenyon, D.M. (2006). *Development and Field Test of ACCESS for ELLs®*. (WIDA Consortium Technical Report No. 1).

Linacre, J.M. (1999). Category Disordering (disordered categories) vs. Threshold Disordering (disordered thresholds). *Rasch Measurement Transactions*, *13*(1), 675.Linacre, J.M. (2002). What Do Infit and Outfit, Mean-Square and Standardized Mean?, *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J.M. (2004). Optimizing Rating Scale Category Effectiveness. In *Introduction to Rasch Measurement, Theory, Models and Applications*. JAM Press: Maple Grove, Minnesota.

Linacre, J.M. (2006). Winsteps (Version 3.60) [computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M., & Wright, B. D. (2006). Winsteps (Version 3.47) [computer software]. Chicago, IL: MESA Press.

Smith R.M. (1996). Polytomous Mean-Square Fit Statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

WIDA Consortium (2012a). *WIDA Alternate ACCESS for ELLs™ Test Administration Manual.* Retrieved from http://www.wida.us/assessment/alternateaccess.aspx.

WIDA Consortium (2012b). *WIDA Alternate ACCESS for ELLs™ Test Administration Tutorial* [online presentation and training videos]. Retrieved from http://www.wida.us/assessment/alternateaccess.aspx.

9 Acknowledgements

The following CAL and WIDA staff contributed to the creation of this document:

Shu Jing Yen, Ph.D.
Deepak Ebenezer
Jennifer Renn, Ph.D.
Melissa Amos
Cathy Cameron
Mohammed Louguit, Ph. D
Carsten Wilmes, Ph.D.
Erin Arango-Escalante
Dorry Kenyon, Ph. D.