World-class Instructional Design and Assessment

Development and Field Test of WIDA MODEL^{TM}

Grades 6–8 and 9–12

Prepared by:

CAL/WIDA Partnership Activities
Psychometrics/Research Team

Center for Applied Linguistics

July 17, 2014

Copyright Notice

2014 WIDA Consortium Members

Alabama
Alaska
Colorado
Delaware
District of Columbia
Georgia
Hawaii
Illinois
Kentucky
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nevada
New Hampshire
New Jersey
New Mexico
North Carolina
North Dakota
Northern Mariana Islands
Oklahoma
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Utah
Vermont
Virginia
Wisconsin
Wyoming

2014 Non-member States Formally Adopting the WIDA ELD Standards

Idaho
Indiana

WIDA™ World-Class Instructional Design and Assessment

WIDA advances academic language development and academic achievement for linguistically diverse students through high quality standards, assessments, research, and professional development for educators. WIDA's vision is to be the most trusted resource in the education of Pre-kindergarten through Grade 12 language learners.

Executive Summary

The World-class Instructional Design and Assessment (WIDA) Measure of Developing English Language (MODEL)™ is an off-the-shelf series of academic English language proficiency assessments for English Language Learners (ELLs) in Kindergarten through Grade 12. The test for Kindergarten was developed from 2006–2008 and became available to WIDA Consortium members and non-members in October 2008. The test for Grades 1–2 and the test for Grades 3–5 were developed from 2008–2010 and became available in August 2010. The test for Grades 6–8 and the test for Grades 9–12 were developed from 2009–2011 and became available in September 2011.

The purpose of this technical report is to describe the development and field test of WIDA MODEL for Grades 6–8 and 9–12. The development and field tests of WIDA MODEL for Kindergarten and Grades 1–2 and 3–5 are discussed in separate technical reports.

This report about WIDA MODEL for Grades 6–8 and 9–12 provides background information about the purposes, format, and scores (Chapter 1); describes how the tests were developed (Chapter 2) and field tested (Chapter 3), presents technical properties of the field tested items and tasks (Chapter 4); explains the linking of WIDA MODEL to the WIDA English Language Proficiency (ELP) levels[1] (Chapter 5); and provides an argument-based framework to support the validity and reliability of the test (Chapter 6).

Summary Highlights

Background Information (Chapter 1)

WIDA MODEL is a series of English language proficiency assessments, which evaluates ELL students' academic English language proficiency in the four language domains of Speaking, Listening, Reading, and Writing. All items and tasks in those sections are aligned to the WIDA ELP Standards (i.e., Social and Instructional Language, Language of Language Arts, Language of Mathematics, Language of Science, and Language of Social Studies). WIDA MODEL can be used to determine the academic English language proficiency level of students who are new to a school or to the U.S. school system and to identify and place students who are candidates for English as a Second Language (ESL) and/or bilingual services. In addition, in states that are members of the WIDA Consortium, WIDA MODEL may be used to determine tier placement on the WIDA ACCESS for ELLs® test (hereafter referred to as ACCESS), to track students' proficiency at an

[1] WIDA MODEL was developed using the 2007 ELP Standards (Gottlieb, Cranley, & Cammilleri, 2007). These standards were updated in the *2012 Amplification of the English Language Development Standards, Kindergarten-Grade 12,* which can be found on the WIDA website (www.wida.us/standards/eld.aspx).

additional time during the school year, and to replace the WIDA-ACCESS Placement Test (W-APT)™.

WIDA MODEL contains both a full-length assessment (MODEL) and a Screener (MODEL Screener), which includes all tasks from the Speaking and Writing sections of MODEL but fewer items in the Listening and Reading sections. MODEL Screener was developed because stakeholders saw a need for a less time-consuming test that would still determine students' language proficiency levels, tier placement on ACCESS, and need for ELL services. MODEL Screener, however, cannot be used to determine amount, type, or exiting of ELL services.

In both MODEL and MODEL Screener, the Speaking section consists of constructed-response tasks that target progressively higher proficiency levels and are administered to individual students in an interview format. The Listening section in MODEL has multiple-choice items, is administered to individual students, and has Low, Mid, and High placement levels so students take only items that are appropriate for their proficiency level. The Reading section in MODEL is also multiple-choice and has Low, Mid, and High placement levels, but the section may be administered to individual students or to a group. The Listening and Reading sections in MODEL Screener have the same format and administration as in MODEL, but they contain fewer items and students are not placed into different levels. The Writing section in both MODEL and MODEL Screener contains two parts: Part A, which asks students to respond to open-ended questions that require only short answers; and Part B, which requires a more extended response that is administered only if students are able to meet expectations on Part A. The Writing section may be administered to individual students or to a group.

After completing a test administration with a student, the test administrator uses lookup tables to convert raw scores to scale scores and proficiency levels. Scores are computed for all four language domains as well as three composite scores—Oral language (Listening and Speaking), Literacy (Reading and Writing), and Overall (all four domains). Proficiency level scores render a student's scale score in terms of the WIDA ELP Standards.

Test Development (Chapter 2)

The WIDA MODEL tests for Grades 6–8 and 9–12 originally used folders of items that were retired or removed from the ACCESS operational test, resulting in a format that is patterned after the ACCESS tests for Grades 6–8 and 9–12. Due to limited availability of retired folders, additional folders were either selected from the ACCESS field test or were newly created by Center for Applied Linguistics (CAL) item writers or external item writers.

All items underwent a series of reviews to ensure that items contained the appropriate content for each grade level and proficiency level, were appropriate and universal to people of different ethnic backgrounds, and did not contain cultural bias or sensitive topics. In addition, cognitive labs were held to collect information about administration procedures and times, accurate placement of students in Low, Mid, or High levels, quality of text and graphics, and the ability of items and tasks to elicit expected language. A number of quality checks, such as proofing and key checks, were conducted before the WIDA MODEL test forms were finalized.

Field Test (Chapter 3)

Field testing for Grades 6–8 and 9–12 was conducted with 1,256 students in 24 schools in four WIDA states—Illinois, Kentucky, Maine, and New Mexico—from November through December 2010. CAL hired field test administrators to assist with the testing of students. The field test administrators followed the same procedures, administration, and scoring guidelines as would be used for operational testing.

Field Test Results (Chapter 4)

Raw data for the Speaking, Listening, and Reading sections were entered and cleaned electronically. The items were scored dichotomously as correct or incorrect so the functioning of items could be analyzed psychometrically and total raw scores could be calculated. An outlier analysis was conducted on the data for the Listening and Reading domains to ensure that students that had discrepant performances on MODEL and ACCESS were not included in calibration and linking procedures. Following the removal of outliers, Rasch analyses revealed that, overall, the Speaking, Listening, and Reading items are productive for measurement and measure the intended construct. For the Writing sections, sets of writing samples were selected to calibrate test development staff at CAL and, later, outside consultants. After all raters were trained to reliably score the calibration samples, all writing samples were rated operationally.

Linking WIDA MODEL to WIDA ELP Levels (Chapter 5)

To facilitate the interpretation of WIDA MODEL scores, scores on the test were linked to scores on ACCESS so they can be understood in terms of the WIDA ELP Levels (Level 1 Entering through Level 5 Bridging). Linking studies were conducted in order to produce lookup tables, which show the proficiency level scores that correspond to students' raw scores and scale scores for each grade, placement level, and domain. In the linking study for Listening and Reading, psychometric methods were used to link WIDA MODEL scores to ACCESS scores. For Writing and Speaking, expert panels qualitatively interpreted performances on WIDA MODEL based on the WIDA ELP levels.

Validity (Chapter 6)

The validity argument presented in this chapter links students' test performance on WIDA MODEL to test scores and provides evidence to support claims related to the quality and consistency of the assessment data gathered during the MODEL field test. Other chapters of the report are referenced in support of the content validity and construct validity of MODEL. In support of the concurrent validity, students' scale scores on MODEL were correlated with their scale scores on ACCESS. Correlations between the Overall Composite scores for MODEL scale scores and the Overall Composite scores for ACCESS were moderate to high.

Table of Contents

1. Background

The World-class Instructional Design and Assessment (WIDA) Measure of Developing English Language (MODEL) ™ is an off-the-shelf series of English language proficiency assessments for Kindergarten through Grade 12. Available to schools around the world, WIDA MODEL assessments can be used by educators to identify newly enrolled students as English Language Learners (ELLs), to place students in ELL services, or to monitor interim progress. WIDA MODEL consists of a comprehensive assessment (hereafter referred to as MODEL) and an abbreviated Screener form (hereafter MODEL Screener). MODEL is a more in-depth and comprehensive test of academic English language proficiency, which can be used for a broader range of purposes than MODEL Screener (as described in Chapter 1.1). MODEL Screener provides an overall proficiency level and can be used as a pre-test or to determine if a student is eligible for ELL services.

WIDA MODEL test items are written from the Model Performance Indicators (MPIs) of WIDA's five English Language Proficiency (ELP) Standards, and each test form assesses the four language domains of Listening, Speaking, Reading, and Writing. WIDA MODEL is an adaptive test that allows flexible placement within sections of the test based on student performance. Test forms for five grade-level clusters have been rolled out incrementally: Kindergarten in October 2008, Grades 1–2 and 3–5 in August 2010, and Grades 6–8 and 9–12 in September 2011.

MODEL Screener was conceptualized in the fall of 2009. Test developers acknowledged that the challenge in developing MODEL Screener was to balance the length of the assessment with the amount of information needed to give a reliable overall proficiency level score.

The rest of this chapter explains WIDA MODEL in more detail.

1.1. Purposes of WIDA MODEL

MODEL can be used for the following purposes:

- To determine the academic English language proficiency level of students who are new to a school or a school system where English is the language of instruction; and
- To identify and place students who are candidates for English as a Second Language (ESL) and/or bilingual services.

In member states of the WIDA Consortium, MODEL may be used for these additional purposes:

- To determine tier placement on ACCESS;
- To track students' proficiency at an additional time during the school year; and

- To replace the WIDA-ACCESS Placement Test (W-APT)™ as the assessment used for program placement of incoming ELL students.

For any of these purposes, scores on MODEL should be considered as only one of several elements in the decision-making process regarding ELL identification and placement in instructional services.

MODEL Screener contains abridged Listening and Reading sections, making it inappropriate for assessing individual domains. MODEL Screener is intended to proffer only an overall assessment of students' English language proficiency and cannot provide guidance on type and amount of ELL services, cannot be used to exit a student from an ELL program, and cannot serve as an interim benchmark assessment. MODEL Screener provides an overall proficiency level score that can be used for identification and placement in ELL services and for determination of tier placement for ACCESS. However, some precision in the measurement of students' English language proficiency is sacrificed as a result of its brevity. For any purpose, scores on MODEL Screener should be considered as only one of several indicators in the decision process regarding ELL services. Table 1 below summarizes the purposes of both MODEL and MODEL Screener.

Table 1

Purposes of MODEL and MODEL Screener

Assessment Purpose	MODEL	MODEL Screener
To determine whether a student needs ELL services	Yes	Yes
To determine English proficiency level on the WIDA scale	Yes	Yes[2]
To provide guidance on the amount and type of ELL services that may be needed	Yes	No
To determine tier placement for ACCESS for ELLs®	Yes	Yes
To exit a student from an ELL program, in conjunction with other evidence	Yes	No
To serve as an interim benchmark assessment	Yes	No

1.2. Underlying Principles of WIDA MODEL

1.2.1. Alignment with the WIDA ELP Standards

WIDA MODEL was developed by the WIDA Consortium and the Center for Applied Linguistics (CAL) as part of a complete system of products and services for K–12 ELLs. From its conceptualization to its launch, WIDA MODEL was planned to be a comprehensive ELP

[2] Although MODEL Screener does provide an English language proficiency level, this determination is based on fewer test items and should be considered as a guideline for proficiency level assignment.

exam assessing students' English language proficiency in the five WIDA ELP Standards[3]. The Standards, their short forms, and abbreviations are:

Standard 1 - English language learners **communicate** for **Social** and **Instructional** purposes within the school setting (**Social and Instructional Language/SIL**);

Standard 2 - English language learners **communicate** information, ideas, and concepts necessary for academic success in the content area of **Language Arts (Language of Language Arts/LoLA)**;

Standard 3 - English language learners **communicate** information, ideas, and concepts necessary for academic success in the content area of **Mathematics (Language of Mathematics/LoMA)**;

Standard 4 - English language learners **communicate** information, ideas, and concepts necessary for academic success in the content area of **Science (Language of Science/LoSC)**; and

Standard 5 - English language learners **communicate** information, ideas, and concepts necessary for academic success in the content area of **Social Studies (Language of Social Studies/LoSS)**.

First published in 2004, the WIDA ELP Standards were developed by WIDA Consortium members with funding from a U.S. Department of Education Enhanced Assessment Grant. The Standards were grounded in scientifically-based research on best practices in general education, English as a Second Language (ESL), and bilingual education. The Standards address the need for students to become fully proficient in both social and academic English. Every selected-response item and every performance-based task on WIDA MODEL targets at least one of these five Standards.

1.2.2. Language Domains

Each of the five WIDA ELP Standards encompasses four language domains that define how ELLs process and use language:

Listening - processing, understanding, interpreting, and evaluating spoken language in a variety of situations;

Speaking - engaging in oral communication in a variety of situations for a variety of purposes and audiences;

[3] WIDA MODEL was developed using the 2007 ELP Standards (Gottlieb, Cranley, & Cammilleri, 2007). These standards were updated in the *2012 Amplification of the English Language Development Standards, Kindergarten-Grade 12*, which can be found on the WIDA website (www.wida.us/standards/eld.aspx).

Reading - processing, understanding, interpreting, and evaluating written language, symbols, and text with understanding and fluency; and

Writing - engaging in written communication in a variety of situations for a variety of purposes and audiences.

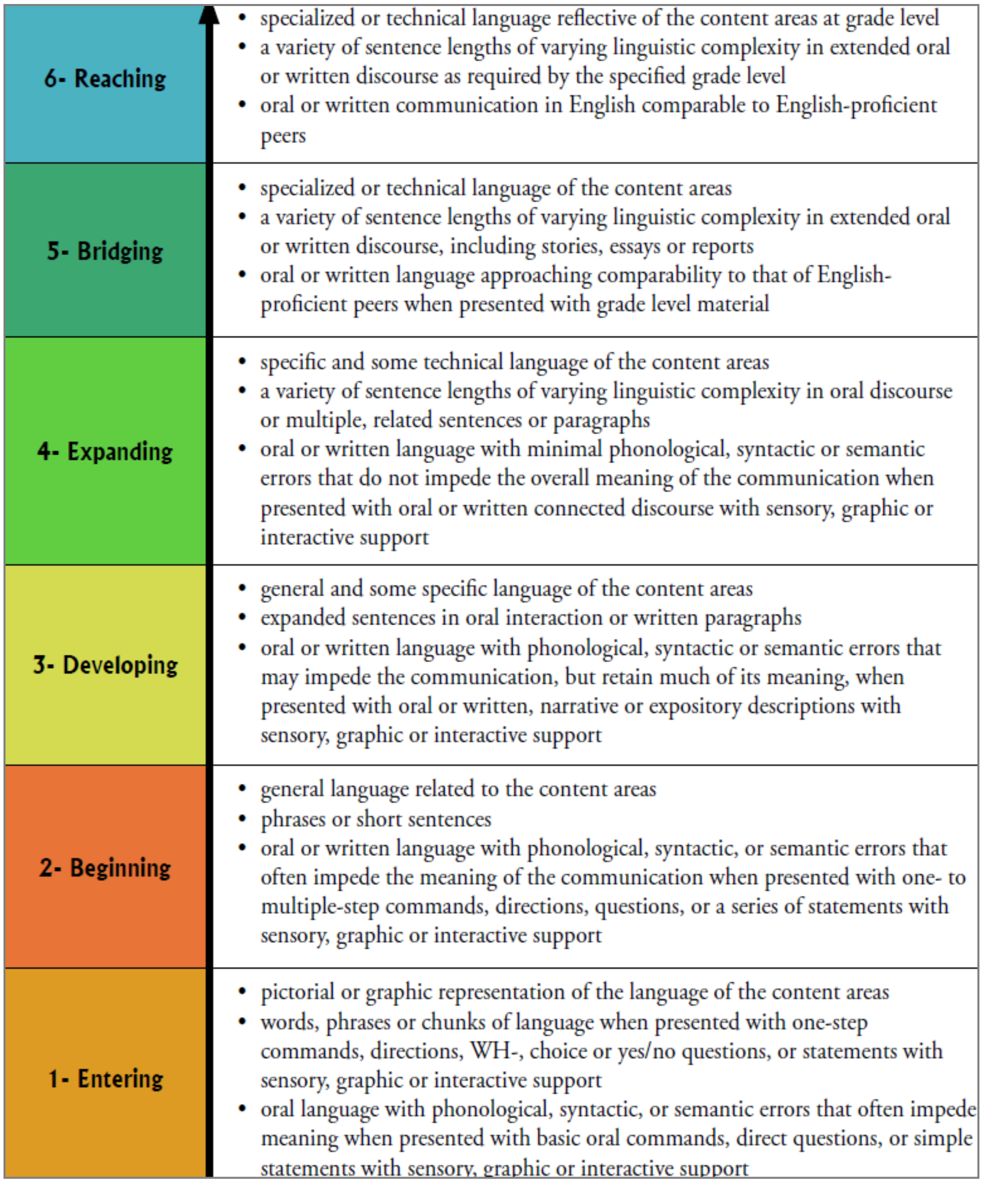
In order to give a full picture of ELL students' English language proficiency, WIDA MODEL assesses proficiency in all four language domains.

1.2.3. Proficiency Levels

The WIDA ELP Standards framework divides the continuum of language development into five proficiency levels: "Entering," "Beginning," "Developing," "Expanding," and "Bridging." The "ceiling" of English language proficiency defined by the Standards for assessment purposes is called "Reaching." The five defined language proficiency levels are embedded in the WIDA ELP Standards in both the Performance Definitions and the MPIs.

1.2.4. Performance Definitions

The WIDA ELP Standards are framed by the WIDA Performance Definitions, which are descriptions of what students are expected to do at each of the six levels of language proficiency. The Performance Definitions are used for interpreting proficiency levels in a general way that is not specific to grade-level clusters or language domains. The definitions were developed using three criteria: 1) Linguistic Complexity: the amount and quality of speech or writing for a given situation; 2) Vocabulary Usage: the specificity of words or phrases for a given context; and 3) Language Control: the comprehensibility of the communication based on the number and type of errors. Figure 1 shows the Performance Definitions for each proficiency level.

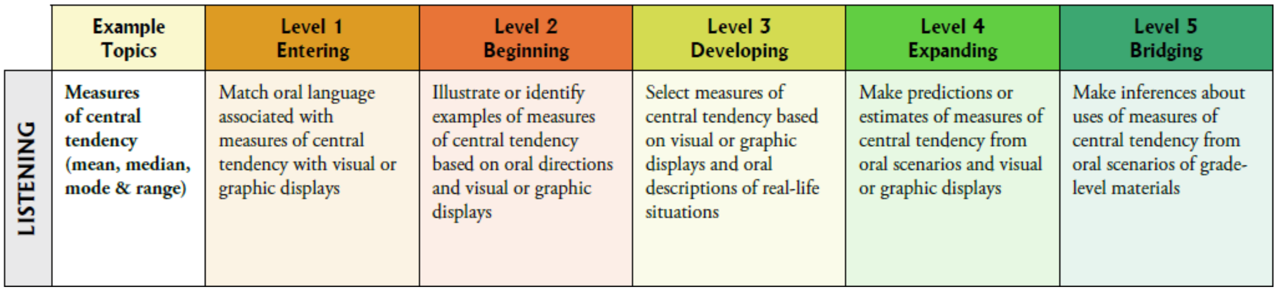


6- Reaching	• specialized or technical language reflective of the content areas at grade level • a variety of sentence lengths of varying linguistic complexity in extended oral or written discourse as required by the specified grade level • oral or written communication in English comparable to English-proficient peers
5- Bridging	• specialized or technical language of the content areas • a variety of sentence lengths of varying linguistic complexity in extended oral or written discourse, including stories, essays or reports • oral or written language approaching comparability to that of English-proficient peers when presented with grade level material
4- Expanding	• specific and some technical language of the content areas • a variety of sentence lengths of varying linguistic complexity in oral discourse or multiple, related sentences or paragraphs • oral or written language with minimal phonological, syntactic or semantic errors that do not impede the overall meaning of the communication when presented with oral or written connected discourse with sensory, graphic or interactive support
3- Developing	• general and some specific language of the content areas • expanded sentences in oral interaction or written paragraphs • oral or written language with phonological, syntactic or semantic errors that may impede the communication, but retain much of its meaning, when presented with oral or written, narrative or expository descriptions with sensory, graphic or interactive support
2- Beginning	• general language related to the content areas • phrases or short sentences • oral or written language with phonological, syntactic, or semantic errors that often impede the meaning of the communication when presented with one- to multiple-step commands, directions, questions, or a series of statements with sensory, graphic or interactive support
1- Entering	• pictorial or graphic representation of the language of the content areas • words, phrases or chunks of language when presented with one-step commands, directions, WH-, choice or yes/no questions, or statements with sensory, graphic or interactive support • oral language with phonological, syntactic, or semantic errors that often impede meaning when presented with basic oral commands, direct questions, or simple statements with sensory, graphic or interactive support

Figure 1: WIDA ELP Levels and Performance Definitions
Source: *Understanding the WIDA English Language Proficiency Standards: A Resource Guide* (Gottlieb, Cranley, & Cammilleri, 2007)

1.2.5. Model Performance Indicators (MPIs)

The WIDA ELP Standards are operationalized into strands of Model Performance Indicators (MPIs), which are the basis for all item specifications for WIDA assessments. MPIs address example topics or genres that have been identified from state academic content standards. Each MPI represents a specific language skill, rather than content or background knowledge. The MPIs give examples of what students should be able to process and produce at a given language proficiency level for a specific grade-level cluster, standard, and domain. Figure 2 shows an example of an MPI for the Language of Mathematics Standard in the WIDA MODEL Listening section for Grades 6–8. This example shows the topic "Measures of Central Tendency" and how middle school students' comprehension progresses as they move through the English Language Proficiency Levels 1–5.



	Example Topics	Level 1 Entering	Level 2 Beginning	Level 3 Developing	Level 4 Expanding	Level 5 Bridging
LISTENING	**Measures of central tendency (mean, median, mode & range)**	Match oral language associated with measures of central tendency with visual or graphic displays	Illustrate or identify examples of measures of central tendency based on oral directions and visual or graphic displays	Select measures of central tendency based on visual or graphic displays and oral descriptions of real-life situations	Make predictions or estimates of measures of central tendency from oral scenarios and visual or graphic displays	Make inferences about uses of measures of central tendency from oral scenarios of grade-level materials

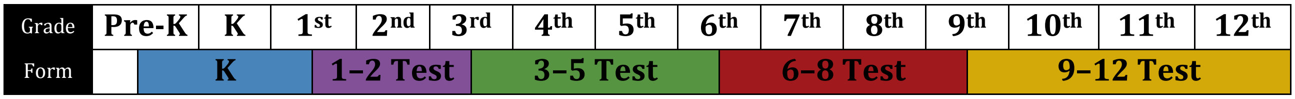
Figure 2: A Strand of MPIs with an Example Topic
Source: *WIDA Consortium English Language Proficiency Standards Grade 6 through Grade 12 2007 Edition* (WIDA, 2007)

1.3. Format of WIDA MODEL

While Chapter 1.2 laid out the organizing principles that underlie WIDA MODEL, Chapter 1.3 describes the structure of the test series.

1.3.1. Grade-level Clusters

WIDA MODEL has test forms for Kindergarten, Grades 1–2, Grades 3–5, Grades 6–8, and Grades 9–12. The appropriate form to administer to a student depends on the current grade of the student and the time of year when the test is administered. Students in the lowest grade in a grade-level cluster should be given the form for the previous grade-level cluster if it is the first semester of the school year. For example, as seen in Figure 3, students in the first semester of sixth grade should take the form for Grades 3–5, and students from the second semester of sixth grade through the first semester of ninth grade should take the form for Grades 6–8. WIDA made these recommendations because students just entering a new grade-level cluster have not yet been exposed to the language proficiency standards and content topics for that cluster.



Grade	Pre-K	K	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
Form	K		1–2 Test		3–5 Test			6–8 Test			9–12 Test			

Figure 3: Appropriate Form of WIDA MODEL Based on Grade Level and Semester
Source: *WIDA MODEL™ Test Administration Manual* (MetriTech & CAL, 2011)

1.3.2. Adaptivity

The assessment is adaptive in order to meet the needs of students at different levels of proficiency. Test items and tasks that allow students at Proficiency Level 1 or 2 to demonstrate the full extent of their language proficiency may not challenge students at Proficiency Level 4 or 5. Likewise, items and tasks developed for students at Proficiency Level 4 or 5 are likely to be far too challenging for students at Proficiency Level 1 or 2.

To match the challenge level of tasks to the proficiency level of the test taker, WIDA MODEL uses adaptive placement in the Listening and Reading sections. A student completes a set of four test items in Step 1 and then takes only selected Step 2 items based on his or her performance on the Step 1 items. A student can be placed into one of three overlapping Step 2 placements: Low, Mid, or High. Each Step 2 placement includes items that assess a range of proficiency levels.

As seen in Figure 4 below, Step 2 Low covers Proficiency Levels 1–3 (Entering through Developing), Step 2 Mid covers Proficiency Levels 2–5 (Beginning through Bridging), and Step 2 High covers Proficiency Levels 3–5 (Developing through Bridging). The High level does not cover Proficiency Level 6, Reaching, because this level represents the end of the continuum rather than another level of language proficiency. The test and the placement rules are designed so that most students will be placed in the Step 2 Mid level. Only students at the very lowest levels of proficiency will be placed in the Step 2 Low level, and only students at the highest levels of proficiency will be placed in the Step 2 High level.

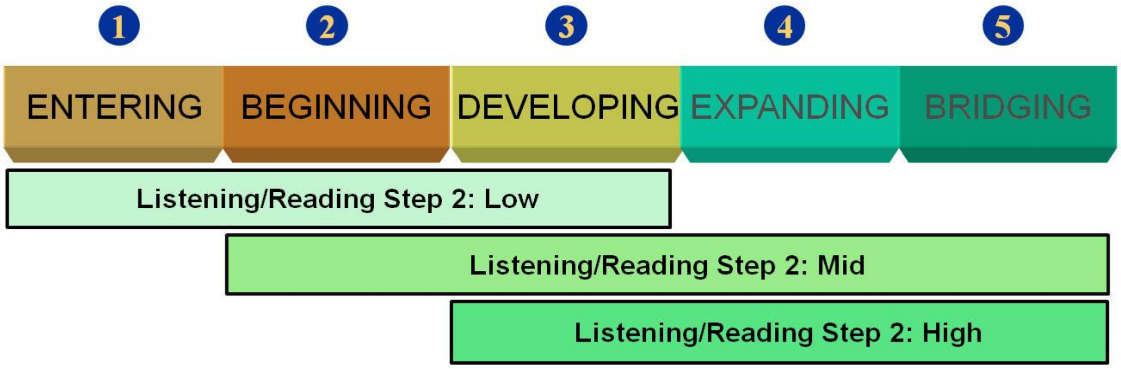

1
2
3
4
5
ENTERING
BEGINNING
DEVELOPING
EXPANDING
BRIDGING
Listening/Reading Step 2: Low
Listening/Reading Step 2: Mid
Listening/Reading Step 2: High

Figure 4: WIDA ELP Levels and WIDA MODEL Step 2 Placement Levels
Source: *WIDA MODEL Test Administration Manual, Grades 6-8* (MetriTech & CAL, 2011)

1.4. Test Administration

This section describes the administration of MODEL and MODEL Screener. The administration procedures for MODEL are explicated in Chapter 1.4.1; procedures for the abbreviated MODEL Screener are outlined in Chapter 1.4.2.

1.4.1. MODEL Administration

MODEL consists of four domain sections: Speaking, Listening, Writing, and Reading. Each domain section is organized into “folders,” or thematic sets of items or tasks with increasing linguistic demand. Figure 5 shows the sequence that test administrators follow to administer the domain sections of MODEL.

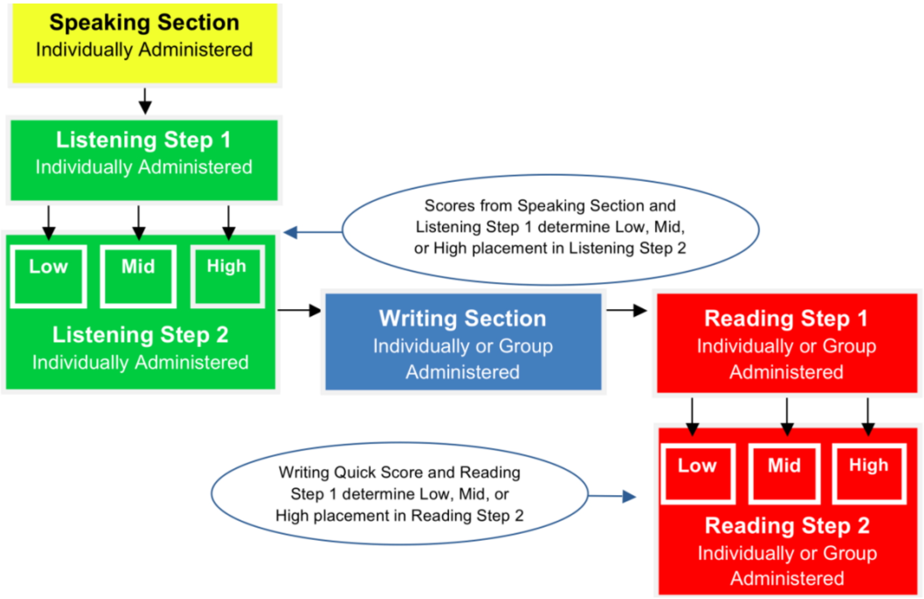

Speaking Section
Individually Administered
Listening Step 1
Individually Administered
Scores from Speaking Section and Listening Step 1 determine Low, Mid, or High placement in Listening Step 2
Low
Mid
High
Listening Step 2
Individually Administered
Writing Section
Individually or Group Administered
Reading Step 1
Individually or Group Administered
Writing Quick Score and Reading Step 1 determine Low, Mid, or High placement in Reading Step 2
Low
Mid
High
Reading Step 2
Individually or Group Administered

Figure 5: Administration Sequence of WIDA MODEL Domains
Source: *WIDA MODEL^{TM} Test Administration Manual, Grades 6-8* (MetriTech & CAL, 2011)

MODEL begins with the Speaking section, which is individually administered to students in an interview format. The Speaking section is comprised of two folders, each with five tasks. These folders address the standards of SIL, LoLA, and LoSS and include tasks targeted at Proficiency Levels 1 through 5. The test administrator asks the student questions targeting progressively higher proficiency levels until the student is no longer able to respond in a way that meets the linguistic demands of the task. When a student's response to a task does not meet expectations, the test administrator stops administering tasks from that folder and either moves on to the next Speaking folder or to the next domain test. Administration of the entire Speaking section lasts approximately fifteen minutes.

The next section, Listening, is also individually administered. It makes use of the adaptive placement described in Chapter 1.3.2. The Listening section consists of a series of passages that are read aloud by the test administrator, followed by multiple-choice questions that are completed by the student. All students complete a set of practice items followed by Listening Step 1, a folder of four items presented in increasing order of linguistic difficulty. For each item, the student either points to his or her answer in the test booklet or says the answer out loud, and the test administrator records the answer in the Student Response Booklet. Then, a placement of Low, Mid, or High is determined in Step 2 using results from the Speaking and Listening Step 1 sections. Step 2 contains 9–12 items, which are grouped into three or four three-item folders depending on the form to which the student is assigned. Administration of the entire Listening section takes approximately 30 minutes.

The Writing section can be administered individually or in a small group. There are two Writing tasks for the TA to choose from; these tasks involve different topics but are meant to elicit the same level of writing. If the Writing section is administered to a small group, the entire group must complete the same task (i.e., Task 1 or Task 2). Each task has its own booklet, and only one booklet is administered to a student at any given time. A test administrator may choose either booklet for a student and may want to use one booklet as an initial assessment tool and the other booklet at a later date to chart growth or collect more information. The tasks have two parts, Part A and Part B, which share a theme. Part A asks students to respond to open-ended questions that require only short answers, and Part B requires a more extended response. A student moves on to Part B only if he or she is able to meet expectations on Part A (see Chapter 3.4.3 for more information about the scoring). When the student has completed the Writing section, the test administrator assigns a Writing Quick Score using scoring criteria in the Student Response Booklet. The Writing Quick Score is based on a reduced version of the WIDA Consortium's Writing Rubric (see Chapter 4.2.1) and is intended to assist with assignment into the appropriate Reading placement level. Administration time for a Writing task is about one minute for Part A and up to 25 minutes for Part B.

The Reading section may be administered individually or continued with the same small group of students as the Writing Section. The Reading section consists of a series of written passages followed by multiple-choice questions. As with the Listening test, each student first completes Step 1, a folder of four items that are progressively more demanding. The test administrator then uses a tally of the number of correct items in Step 1 and the Writing Quick Score to assign the student to the appropriate Reading Step 2 placement of Low, Mid, or High. Step 2 contains 9–12 items from three or four three-item folders, depending on the form to which the student is assigned. Students record their answers by bubbling them in the Student Response Booklet. The Reading section is designed to take up 20–25 minutes depending on the placement level.

1.4.2. MODEL Screener Administration

MODEL Screener is an abbreviated version of MODEL that can be used to determine if a student is eligible for ELL services. MODEL Screener includes the same Speaking and Writing tasks as MODEL but has fewer Listening and Reading items. The Listening and Reading sections of MODEL Screener consist of the same four-item folder in Step 1 of MODEL but have only one additional four-item folder in Step 2. MODEL Screener uses the same set of materials as MODEL, and within the MODEL test materials, any folders that are also used for administering MODEL Screener are denoted with an orange bar on the page.

Figure 6 highlights the domain parts and steps that are used for MODEL Screener. (Note that, in this context, the letters A, B, C, and D refer to folders, not tiers.) When administering MODEL Screener, a test administrator first administers and scores Part A and Part B of the Speaking section. The test administrator next administers and scores the Listening Practice and Listening Step 1. Applying placement rules to these scores, the test administrator determines whether the student has an adequate proficiency level to take Screener Listening Step 2. Students exhibiting very low levels of language proficiency immediately move on to the next domain section of the test after completing Listening Step 1, while students at higher proficiency levels complete the second Listening folder. After finishing the Listening section, the Writing test is administered and scored. The test administrator administers Writing Part A and, if the student is eligible, Writing Part B. Next, the test administrator administers Reading Step 1, after which he or she determines the student's Writing Quick Score (Low, Mid, or High). Then, following the criteria given in the Screener Reading Placement, the test administrator determines whether the student has an adequate proficiency level to take the Screener Reading Step 2. Students who perform at higher levels take the second folder, while students who do not exhibit adequate proficiency stop the Reading portion after completing Step 1.

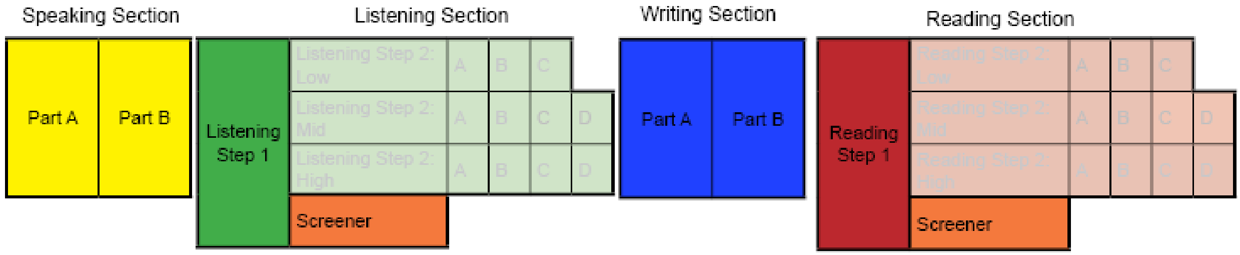


Figure 6: Components of MODEL Screener
Source: *WIDA MODEL^{TM} Test Administration Manual, Grades 6-8* (MetriTech & CAL, 2011)

Table 1, presented in Chapter 1.1, recommends which assessment—MODEL or MODEL Screener—to administer to a student based on the intended purpose. Both MODEL and MODEL Screener can be used to determine a student's need for ELL services, his or her overall level of English language proficiency, and his or her tier placement on the ACCESS test. However, the abridged nature of the Listening and Reading portions of MODEL Screener makes it inappropriate for assessing individual domains, particularly at lower levels of proficiency; instead, it is intended to proffer only an overall assessment of student's English language proficiency and cannot provide guidance on type and amount of ELL services, cannot be used to exit a student from an ELL program, and cannot serve as an interim benchmark assessment.

1.5. Test Scores

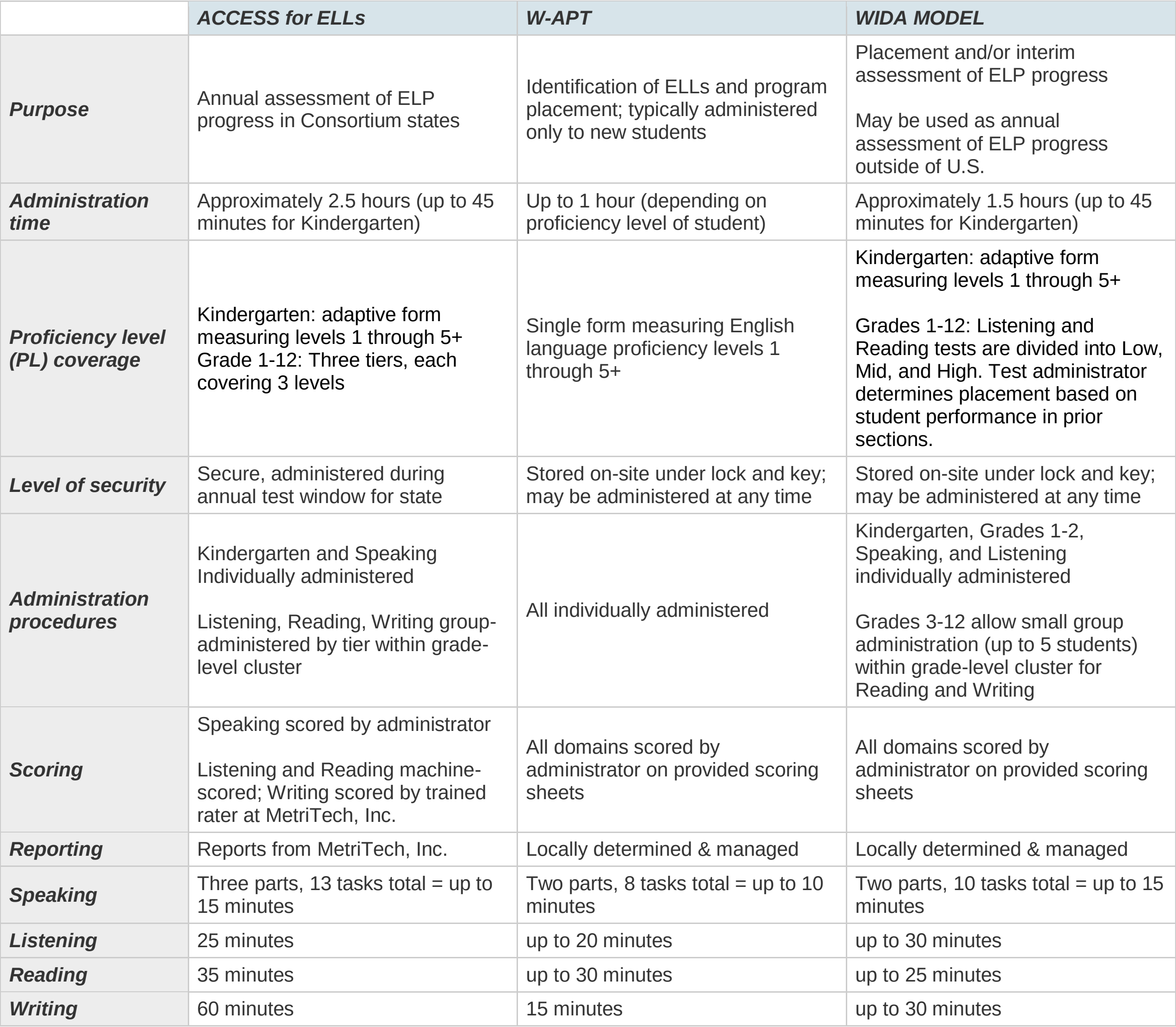
WIDA MODEL scores are reported as both scale scores and proficiency level scores. Scale scores are conversions of raw scores to a common scale that is familiar to test users, that is constant across test forms and grade-level clusters, and that allows comparison among students. Since WIDA MODEL and ACCESS were developed using the same standards and because a reporting scale had been developed and validated for ACCESS (Kenyon, 2006), WIDA MODEL scale scores are reported on the same vertical scale as ACCESS (see Chapter 5 for more information on the linking studies). WIDA MODEL scale scores range from 100 to 600 for all domains and composites.

Proficiency level scores are interpretations of a student’s scale score in terms of the WIDA ELP Standards. These scores range from one to six and are estimated to a tenth of a point. For example, a score of 4.5 indicates that the student’s scale score is halfway between the cut for Proficiency Levels 4 and 5. Because the width between cut scores varies, proficiency level cut scores should not be considered to form an interval scale across proficiency levels.

In addition to the four domains, proficiency level scores are provided for three composite scores: Oral (50% Listening + 50% Speaking), Literacy (50% Reading + 50% Writing), and Overall (35% Reading + 35% Writing + 15% Listening + 15% Speaking). Because the Overall Composite score is based on students’ performances in all four domains, this scale score is recommended as the best choice for use in making educational decisions regarding students’ English language proficiency.

1.6. WIDA MODEL and ACCESS

Users of WIDA MODEL who are already familiar with ACCESS or W-APT may find it helpful to see the related assessments explicitly compared, as in Figure 7 below.



	ACCESS for ELLs	*W-APT*	*WIDA MODEL*
Purpose	Annual assessment of ELP progress in Consortium states	Identification of ELLs and program placement; typically administered only to new students	Placement and/or interim assessment of ELP progress May be used as annual assessment of ELP progress outside of U.S.
Administration time	Approximately 2.5 hours (up to 45 minutes for Kindergarten)	Up to 1 hour (depending on proficiency level of student)	Approximately 1.5 hours (up to 45 minutes for Kindergarten)
Proficiency level (PL) coverage	Kindergarten: adaptive form measuring levels 1 through 5+ Grade 1-12: Three tiers, each covering 3 levels	Single form measuring English language proficiency levels 1 through 5+	Kindergarten: adaptive form measuring levels 1 through 5+ Grades 1-12: Listening and Reading tests are divided into Low, Mid, and High. Test administrator determines placement based on student performance in prior sections.
Level of security	Secure, administered during annual test window for state	Stored on-site under lock and key; may be administered at any time	Stored on-site under lock and key; may be administered at any time
Administration procedures	Kindergarten and Speaking Individually administered Listening, Reading, Writing group-administered by tier within grade-level cluster	All individually administered	Kindergarten, Grades 1-2, Speaking, and Listening individually administered Grades 3-12 allow small group administration (up to 5 students) within grade-level cluster for Reading and Writing
Scoring	Speaking scored by administrator Listening and Reading machine-scored; Writing scored by trained rater at MetriTech, Inc.	All domains scored by administrator on provided scoring sheets	All domains scored by administrator on provided scoring sheets
Reporting	Reports from MetriTech, Inc.	Locally determined & managed	Locally determined & managed
Speaking	Three parts, 13 tasks total = up to 15 minutes	Two parts, 8 tasks total = up to 10 minutes	Two parts, 10 tasks total = up to 15 minutes
Listening	25 minutes	up to 20 minutes	up to 30 minutes
Reading	35 minutes	up to 30 minutes	up to 25 minutes
Writing	60 minutes	15 minutes	up to 30 minutes

Figure 7: Differences among WIDA MODEL, ACCESS for ELLs, and W-APT
Source: Adapted from *Comparing WIDA MODEL™, ACCESS for ELLs*®*, and W-APT*™ (WIDA, 2011a)

2. Test Development

This chapter details the test development procedures for the WIDA MODEL series for Grades 6-8 and 9-12.

2.1. Test Maps

During the planning stages of the test development in the summer of 2009, CAL managers, with stakeholders' input, created a test map for Grades 6–8 and Grades 9–12 to show which language domains, WIDA ELP Standards, and proficiency levels would be covered in each test form. Additionally, as described in Chapter 1.3.2, it was important that the tests be tailored to student ability with the use of steps and placement levels, similar to the way that ACCESS uses tiers. The test map for Grades 6–8 and 9–12 is shown in Table 2. Note that the test for each grade-level cluster was designed to have the same number of folders for each WIDA ELP Standard and tier.

Table 2

Test Map for WIDA MODEL for Grades 6–8 and Grades 9–12

	Listening		Reading		Writing	Speaking
Test Step and Placement Level	WIDA ELP Standard[4]	Folder Tier[5]	WIDA ELP Standard	Folder Tier	WIDA ELP Standard	WIDA ELP Standard
Step 1	LoLA	B+	LoLA	B+	IT and IT	SIL and LoLA/LoSS
Step 2: Low	SIL	A	LoLA	A		
	LoLA	A	LoMA	A		
	LoMA	A	LoSC	A		
Step 2: Mid	LoMA	B	LoMA	B		
	LoSC	B	LoSC	B		
	LoSS	C	LoSS	B		
	LoLA	C	LoLA	C		
Step 2: High	LoMA	C	LoMA	C		
	LoSC	C	LoSC	C		
	LoSS	C	LoSS	C		
	LoLA	C	LoLA	C		
Step 2: Screener	LoMA/LoSC	C	LoMA/LoSC	C		

[4] As described in Chapter 1.2.1, the acronyms for the WIDA ELP Standards are "SIL" for Social and Instructional Language, "LoLA" for Language of Language Arts, "LoMA" for Language of Mathematics, "LoSC" for Language of Science, and "LoSS" for Language of Social Studies. "IT" indicates an integrated Writing task that includes SIL, LoLA, and LoSS.

[5] "Folder Tier" indicates the proficiency level of items in a folder. Tier A folders have items at Proficiency Levels 1 (Entering), 2 (Beginning), and 3 (Developing). Tier B folders have items at Proficiency Levels 2, 3, and 4 (Expanding). Tier C folders have items at Proficiency Levels 3, 4, and 5 (Bridging). Tier B+ folders have items at Proficiency Levels 2, 3, 4, and 5. Speaking and Writing have no tiers, as they are comprised of tasks that are progressively more demanding and are intended for students of all proficiency levels.

Folders for WIDA MODEL were selected from various sources to meet the test map criteria. First, CAL identified several folders that had been retired, or removed, from operational ACCESS tests. Then, CAL identified several folders that had been field tested for ACCESS Series 201 (2009-2010 academic year) but not used operationally. CAL and WIDA reviewed these retired and field tested folders, cut folders that were not deemed high quality, and saw which gaps in the test map remained to be filled. CAL attempted to fill in some of the gaps by revising certain folders that they had cut and by considering additional retired and field tested ACCESS folders.

After these series of reviews and revisions, a final list of retired or field tested ACCESS folders was selected to appear on the final WIDA MODEL test forms. For Grades 6–8, two Speaking, eight Listening, two Writing, and six Reading retired or field tested ACCESS folders were selected. For Grades 9–12, two Speaking, four Listening, two Writing, and seven Reading retired or field tested ACCESS folders were selected. These folders received new graphics and other necessary revisions. Remaining gaps in the WIDA MODEL test map were filled by folders that were newly developed in-house by CAL staff or by external item writers (see Chapter 2.2 below for details).

2.2. Item Writing

In August 2009, CAL test developers hired external consultants to write items to fill in gaps in the test map. As seen in Table 3, three external item writers developed Reading, Listening, and Speaking[6] folders for Grades 6–8 and 9–12. These item writers had previous experience writing items for ACCESS.

Table 3

Item Writers and Their Affiliations and Assignments

Name	Affiliation	Assignment for Writing Items
Raymond Devenney	Bell Multicultural High School, Washington, DC	6-8 Reading, 9-12 Listening, 9-12 Speaking
Katie Cobb	Charlotte-Mecklenburg Schools, Charlotte, NC	6-8 Reading, 9-12 Listening, 9-12 Reading
Kate Jerris	Independent consultant, NJ	6-8 Listening, 6-8 Reading, 9-12 Listening, 9-12 Reading

The external item writers were given the grade-level clusters, language domains, WIDA ELP Standards, and proficiency levels or tiers for which they would be writing items. Using resources such as textbooks, worksheets, and websites, item writers conceptualized topics from which to build age- and proficiency level-appropriate folders. For each newly developed folder, the item

[6] At the beginning of test development, test developers needed one Speaking folder for Grades 9–12. This Speaking folder was developed by the item writers but was later replaced with a Speaking folder that was retired from ACCESS.

writers filled in a Theme Folder Worksheet. The worksheets contained a page for the folder theme (e.g., theme ID, orientation, theme passage/prompt, theme graphic, and description), a separate page for each item (e.g., item number, proficiency level, item passage/prompt, item graphic, task statement/question, key, distractor 1, distractor 2, and distractor 3), and a page for the sources and notes. After completing the worksheets, the item writers copied and pasted the information from the worksheets into the Online Item Writer, WIDA's online database for uploading and receiving feedback on item writing assignments.

As seen in Table 4, the external item writers drafted two Listening and five Reading folders for Grades 6–8 and seven Listening, two Reading, and one Speaking folders for Grades 9–12. The folders were written to specifications for Tier A, Tier B, and Tier C and addressed the five WIDA ELP Standards (i.e., SIL, LoLA, LoMA, LoSC, and LoSS).

Table 4
Folders Created by the External Item Writers

Grade-level Cluster	Domain	WIDA ELP Standard	Folder Tier	Folder Title
6-8	Listening	LoLA	C	Zebra
6-8	Listening	SIL	A	Travel Advertisement (Poster Project)
6-8	Reading	LoLA	A	Making Cookies (Cooking Eggs)
6-8	Reading	LoMA/LoSC	C	Chesapeake Bay
6-8	Reading	LoSC	A	The Effects of Adrenaline
6-8	Reading	LoSC	A	Convection Currents
6-8	Reading	LoSS	B	Photography Firsts (Industrial Revolution)
9-12	Listening	LoLA	C	Clever Dog and Bow-legged Admiral (Sea Story)
9-12	Listening	LoLA	A	Oranges (One Day After School)
9-12	Listening	LoLA	B/C	Making Bread
9-12	Listening	LoMA	B	Quilt
9-12	Listening	LoMA/LoSC	C	Balance
9-12	Listening	LoSC	B	Blue Crabs
9-12	Listening	LoSS	B/C	Persuasive Techniques
9-12	Reading	LoMA/LoSC	C	Science Experiment (Electrical Circuit)
9-12	Reading	LoSC	B	Model Rockets
9-12	Speaking	LoSS	--	Time

These folders were reviewed during an internal CAL "triage" to see if they should be further developed or should be discarded based on the test map, alignment with standards, desired proficiency level, and range of content.

While the external item writers were writing new items, CAL test developers also wrote new items in order to ensure that gaps in the test maps were filled. Table 5 lists the internal item writers' folders that appeared on the final WIDA MODEL test forms.

Table 5

Folders Created by the Internal Item Writers

Grade-level Cluster	Domain	WIDA ELP Standard	Folder Tier	Folder Title
6-8	Listening	LoLA	C	The Hungry Coat
6-8	Listening	LoSS	C	Renewable Energy
6-8	Listening	LoMA/LoSC	C	Hanging Scale
6-8	Reading	LoSC	C	Pancakes
9-12	Listening	LoLA/LoSS	B+	Group Behavior
9-12	Listening	LoMA	A	Camilla's Plant
9-12	Listening	LoMA	C	Statistics
9-12	Listening	LoSC	C	Single-celled Organisms
9-12	Reading	LoLA	A	Julia Child
9-12	Reading	LoMA	C	Perspective
9-12	Reading	LoSC	C	Bacterial Growth

In October 2009, all drafted folders were sent to WIDA for their review. Refinement of folders continued through 2010. The final list of folders that were used in MODEL is listed in Chapter 2.8 of this report, and the final list of folders included in MODEL Screener is listed in Chapter 2.9.

2.3. Content Review

CAL held a Content Review for the Listening and Reading folders in April and May 2010. The purpose of the Content Review was to ensure that the content of the folders was accessible and relevant to the students in the grade level being assessed.

For the Content Review, CAL recruited one internal employee and three consultants to serve as standards experts for Grades 6–8 and three consultants to serve as standards experts for Grades 9–12. The standards experts' names, affiliations, and assignments are listed in Table 6. These experts were selected because of their teaching experience and content knowledge in math, science, or social studies.

Table 6

Standards Experts and Their Affiliations and Assignments

Name	Affiliation	Assignment for Reviewing Items
Jennifer Lamason	Good Hope Middle School, PA	6-8, LoMA
Dana Mehaffie	East Pennsboro Middle School, PA	6-8, LoSC
Barbara Jean	Daniels Middle School, NC	6-8, LoSS
Elizabeth Oestreich	Wausau East High School, WI	9-12, LoMA
Joe Hushek	Selma High School, CA	9-12, LoSC
Jennifer Himmel	Center for Applied Linguistics, DC	6-8, LoSC
John Spicer	Hibriten High School, NC	9-12, LoSS

Prior to the content review, the standards experts received training, including information about the background and purposes of WIDA MODEL, what a Content Review is and why it is done, how to give the best feedback about folders, and how to ensure confidentiality and security of test folders.

The standards experts participated in the content review remotely rather than traveling to CAL. They each were mailed a packet of folders and were emailed a Content Review Form in Adobe PDF. They had approximately two weeks to review their assigned folders and to complete a review form for each folder.

In order to review the folders, the standards experts were provided with the following instructions:

1. Note the folder topic (provided) and theme (usually the title).
2. Read through the folder once to get an idea of how the theme is presented.
3. Please comment if there is anything factually inaccurate.
4. Have we presented the topic and theme in a way that is appropriate and accessible to students in this grade-level cluster?
    - If the topic and theme are not appropriate for students in that grade cluster, answer the questions/complete the steps below:
        - What is inappropriate about this folder?
        - What can we change about the graphics and text to improve it?
        - What do we need to remove?
        - What do we need to add?
    - If the topic and theme are appropriate and you feel they will be okay to present to students in that grade cluster, answer the questions/complete the steps below:
        - Is the information presented in the items presented in a typical way?
        - What would you change about the items to make them more appropriate for students in that grade cluster?
        - Is there anything we need to remove?

- Is there anything we need to add?

5. What vocabulary do you expose students to when discussing the given topic of the folder? Please give examples of words you would use for the simplest of explanations to the most technical.

After completion of the Content Review, the subject experts returned their paper folders to CAL via FedEx and the review forms to CAL via email.

The subject reviewers' suggestions for the test developers mostly involved specific ways to revise the text of the passages and questions or options to make them more grade-level appropriate or to resemble text that their students see in the classroom. They also made some suggestions to revise graphics. The test developers used this feedback to improve the folders before the field test.

2.4. Bias and Content Review

On November 16, 2009, CAL test developers held a Bias and Content Review for Speaking, Listening, Writing, and Reading folders for Grades 6–8 and 9–12. The review was held at CAL. CAL recruited current or former ESL teachers and current ESL graduate students who had teaching experience with diverse students in Grades 6–12. The goals of the meeting were to identify any bias, inappropriate content, and sensitive issues in the folders and to make suggestions on how to revise the folders.

As shown in Table 7, five consultants participated in the bias and content review for Grades 6–8, and four consultants participated for Grades 9–12.

Table 7
Bias and Content Reviewers

Grade-level Cluster	Bias and Content Reviewers
6-8	Carla Williams, Meryl B. Brady, Joanne Chen, Micki Suchenski, and Marcella Gillis
9-12	Jessica Lopez, Connie Thibeault, Rae Roberts, and Marsha Sprague

At the beginning of the Bias and Content Review, test developers presented a slideshow presentation in order to inform the reviewers about the WIDA Consortium, WIDA MODEL, and the procedures for reviewing bias and content.

The reviewers then formed groups by grade-level cluster and began to review their folders. A facilitator guided the groups' discussions. Each panelist took notes in his or her binder, and a notetaker took notes for the group in Microsoft Excel spreadsheets. The panelists spent approximately 10–15 minutes reviewing each folder to identify problems and to discuss solutions.

To identify possible problems, the panelists referred to the Speaking and Writing rubrics, a Bias and Content Review Checklist, and a Sensitive Topics List. The panelists reviewed the folders to ensure that the content was measuring language at the intended proficiency level, that it was accurate and appropriate for the grade-level cluster, and that it was comprehensible. Then, the panelists reviewed the folders for bias, which would result in differential performance in people who have the same English ability but are from different subgroups (e.g., gender, race/ethnicity, disability, home language, religion, culture, region, and socio-economic status). Examples of bias include stereotypes of subgroups and Spanish-English cognates. Finally, the panelists reviewed the folders for sensitive issues that might elicit strong emotions among test takers and, as a result, prevent those students from accurately demonstrating their academic English proficiency. Examples of sensitive issues that the panelists looked for were violence and height/weight.

Participants suggested revising folders by labeling or reordering a graphic, adding or modifying a sentence in a passage, making people's names less American, and replacing objects and places that are not found in other countries (e.g., radiator, bistro).

2.5. Bias and Sensitivity Review

On May 7, 2010, the test developers held a Bias and Sensitivity Review at CAL. The consultants from the Bias and Content Review in November 2009 (see Chapter 2.4) were invited back to participate. This Bias and Sensitivity Review differed slightly from the Bias and Content Review, as this review focused on newly developed folders rather than previously developed folders, and it more heavily emphasized detecting sensitive topics.

At the beginning of the meeting, test developers provided a presentation to train the panelists. The presentation contained information about the uses of WIDA MODEL, updates about the design of the test forms, and examples of bias and sensitive issues.

The panelists were given five new folders for Grades 6–8 and four new folders for Grades 9–12. The panelists were to spend about 10 minutes reviewing each of the folders for sensitive topics, such as violent activities, religion, gambling, sexuality, war, poverty, disease, death, and prehistoric times. The panelists discussed possible revisions to improve the folders. After reaching a consensus about the problems and revisions for folders, they recorded notes in Microsoft Excel spreadsheets. At the end of the meeting, they submitted these notes to the test developers for consideration in revising folders. Not many biased or sensitive issues were found.

Suggestions included labeling a map, making wording in a passage and question clearer, and simplifying people’s names and technical vocabulary.

2.6. International Perspectives Panel

On May 7, 2010, CAL conducted an International Perspectives Panel on all WIDA MODEL Listening and Reading folders for Grades 6–8 and 9–12. The goal of the panel was to minimize construct-irrelevant content for international students who are not familiar with American contexts (e.g., foods and activities).

As seen in Table 8, participants were seven CAL employees representing different nationalities. These participants had been educated in schools outside of the United States when they were children.

Table 8
Participants of the International Perspectives Panel and Their Country of Origin

Name	Country of Origin
Anna Todorova	Bulgaria
Elizabeth Castillo	Dominican Republic
Marcos Carvalho	Brazil
Mohammed Louguit	Morocco
Olesya Warner	Russia
Rafael Michelena	Venezuela
Xiaomin Huang	China

The folders that were reviewed were newly developed folders as well as folders that had problems during previous reviews. The panel participants reviewed eight folders for Grades 6–8 and five folders for Grades 9–12. The panel review began with the test developers training the reviewers on the background of WIDA MODEL, the purpose of an International Perspectives Panel, and the procedures that they would follow. Then, the reviewers individually followed a protocol to review the folders. The reviewers took notes about the text, questions, images, and their general impressions of the folders. They looked for things that would be unusual, confusing, or jarring to a student taking WIDA MODEL in an international school and would prevent the student from answering the question. Next, the reviewers discussed their opinions with each other, and they generated suggestions for improving the folders. At the end of the workshop, the reviewers debriefed with a test developer. The entire review lasted about two hours.

The reviewers provided suggestions for improving about two-thirds of the items. Some examples of suggestions were replacing American English vocabulary with International English vocabulary (e.g., “fiddle” changed to “violin”), using the metric system rather than United States customary units, using consistent terminology throughout a passage, making a graphic look realistic, and changing an activity to one that is more common in other countries.

2.7. Pilot Testing

Pilot testing generates exploratory, qualitative findings that help test developers to identify necessary revisions for an assessment and to support the validity of the assessment prior to operational use. Various data collection methods (e.g., questionnaires, verbal protocols, observations, interviews, analyses of assessment records, etc.) can be used to gather data about an assessment. Depending on the results of the pilot test, the test developers can decide to retain, modify, or remove parts of the assessment and procedures before field testing occurs.

For WIDA MODEL 6-8 and 9-12, three rounds of pilot testing were conducted for Listening and Reading items, and two rounds of pilot testing were conducted for Speaking tasks. The Writing tasks were not pilot tested because they had been pilot tested for past operational ACCESS tests.

2.7.1. Listening and Reading

During development of the Listening and Reading tests, employees at CAL conducted three rounds of pilot testing: 1) a pilot of cognitive lab procedures and protocols, 2) Cognitive Lab 1 of test items, and 3) Cognitive Lab 2 of test items. For the pilot and two cognitive labs, CAL recruited schools in the Washington, DC area via phone calls and emails, offering monetary and gift incentives (e.g., pizza).

The pilot was used to confirm that processes and procedures would elicit the necessary feedback to improve item quality, and the cognitive labs gathered qualitative data from students on the functioning of WIDA MODEL, including information about the script, items and tasks, and student performances, in order to revise test items prior to field testing. The details of each of these steps are detailed in the following sections.

2.7.1.1. Pilot Test

On June 14–16, 2010, five CAL and WIDA employees conducted a pilot of cognitive lab procedures and protocols at Martin Luther King, Jr. Middle School in Beltsville, MD. The pilot included video recorded interviews with 8 seventh graders and 14 eighth graders of varying English language proficiency levels, genders, and native languages.

The main purpose of the pilot test was to confirm that the interview protocol could elicit necessary information from students about item performance. In advance of the pilot test, researchers gathered several retired ACCESS Listening and Reading items for which quantitative results had already been obtained on past operational ACCESS tests (e.g., Series 101 in 2005-2006 and Series 102 in 2006-2007). These items, which were not intended to be included in WIDA MODEL, were used to determine if the interview protocol provided complete and accurate information about item functionality; that is, the items were used to ascertain whether the qualitative feedback from students accurately supported the psychometric statistics. If the protocol did not elicit this information, test developers were able to refine the protocol. Additionally, the procedures for the cognitive labs were piloted to ensure that the logistics of scheduling, note-taking, etc. went smoothly.

The interview with each student lasted between 45 and 60 minutes. The interview protocol started with the interviewer and notetaker introducing themselves to the student and asking in which language the interview should be conducted. Spanish was used with students of low English proficiency, and English was used with students of higher English proficiency. Test developers also asked if the student minded being filmed and assured each student that his or her face would not be shown and that the video would be used only for research purposes. The camera was turned on only if the student provided permission to do so. The interviewer then explained what the test is and the purpose of doing the interview, followed by a demonstration with the notetaker to show how the interview would be conducted. After this demonstration, the interviewer confirmed that the student understood the process and that interviewing could begin. Depending on which domain the interviewer had selected for the student beforehand, the student listened to an audio recording of Listening items or silently read Reading items from a booklet; then the student marked his or her responses on a separate answer sheet.

The interviewer had each student participate in a discussion of the student's answers at the end of each folder. The notetaker took detailed notes on the student's responses to interview questions like the following:

- Was there anything confusing about the question or the answers? Was there anything that you did not understand?
- Do you think any other answer could be correct? Why?
- Do you know what this specific vocabulary word means? Do you know the kind of pictures that were used in the answer choices?
- Have you studied this kind of subject in school? What did you learn?
- What did you think of the test overall? Have you ever taken a test like this before? Do you have any suggestions for us?
- Please read the question out loud to me. What does this question mean in your own words?
- How did you decide which answer to choose? Were there words that helped you? Did you guess? Did you know some answers were not correct? Did you take notes? On what?
- What answer did you choose? What does that answer mean in your own words? How sure are you that your answer is correct?
- Please explain what the passage was about. Was there anything that you didn't understand, or that you thought was too difficult?

- What do the pictures show you? Did the picture help you to answer the questions? Was there anything confusing about the picture?

Responses were recorded in observation protocol documents, which were aligned with the questions being asked in the interview protocol. This form had columns for the student's responses and rows for the topics and questions that the interviewer asked.

After completion of test administration, two pairs of two raters watched the videos of the interviews and reviewed the interview notes. They coded the students' responses in the notes with the following codes:

A. Student admittedly did not understand question or response options and guessed.
B. Student misunderstood key language (as indicated in his/her paraphrase of questions and/or answer).
C. Information in item was misleading or distracting in some way (student had a reasoned explanation that was a plausible answer).
D. Student background knowledge interfered (e.g., "It's 85 degrees today and you are going to the park. Will you take a coat with you?" Student responds, "Yes because my mother always makes me take my coat").
E. The cognitive task that the student performed was different than the item writer intended and therefore the student did not need to understand construct relevant language (e.g., student used process of elimination instead of synthesis, identified the same word in the reading passage and response options).
F. The graphic or picture was misleading or distracting in some way.
G. Another item in the folder influenced the student's answer.
H. Unknown
I. None of the above (please provide explanation).

Items were analyzed across students within and between proficiency levels, with the intent of identifying which items performed as intended and which items were misperforming. For each item, test developers compared the *a priori* proficiency level for the item's MPI with the empirical proficiency level from operational ACCESS to determine if the item performed as expected. Then, test developers used the notes and codes from the interviews with the pilot test students to select the code that explained why the item did or did not perform as expected. When the intended difficulty of an item matched its observed difficulty, researchers confirmed that students answered based on construct-relevant features of the item. When the intended difficulty of an item did not match its observed difficulty, researchers identified why the difficulties did not match.

After the test developers completed their analyses, they reviewed their findings from the pilot test. They found that the scheduling, length of time, introductions, and note-taking in the pilot worked well. The observation protocols were easy to follow and allowed the notetakers the space needed to take notes. Their main finding from the pilot test was that the majority of questions in the semi-retrospective post-assessment interview protocol elicited useful information from students about the functioning of test items. A few questions, however, did not evoke the intended information. The questions "Was there anything confusing about the question or the answers?" and "Was there anything you did not understand?" did not motivate a significant student response. The question "How did you decide which answer to choose?" was understood as "How did you know that (student answer) was correct, and not (the other three answer choices)?" The question "How sure are you that your answer is correct?" was understood by lower-proficiency students to mean "How did you find your answer?" When students did indicate their surety, they sometimes provided a percentage that seemed arbitrary. These few questions were then revised prior to the subsequent cognitive labs.

2.7.1.2. Cognitive Lab 1

From July 19, 2010 through August 13, 2010, CAL employees and consultants conducted Cognitive Lab 1. The cognitive lab included 37 students in Grades 6–7 and 54 students in Grades 9–12. These students were enrolled in summer school at Hyattsville Middle School in Hyattsville, MD, Thomas Johnson Middle School in Lanham, MD, Northwestern High School in Hyattsville, MD, and Osbourn Park High School in Manassas, VA.

Cognitive Lab 1 was exploratory and aimed at determining whether WIDA MODEL Listening and Reading items performed as expected (e.g., if students got the right answers for the right reasons, the right answers for the wrong reasons, the wrong answers for the wrong reasons, etc.) and, if not, how items should be improved so that what they test is construct-relevant. In preparation for the cognitive lab, experienced researchers and new consultants received training on interviewing, video recording, and note-taking. The new consultants were high-proficiency Spanish speakers who would interview students who were native Spanish speakers with low English proficiency. Because the notes taken during the pilot provided the most useful information, only a few students were video recorded in Cognitive Lab 1. The interview and observation protocols in Cognitive Lab 1 were similar to those used in the pilot test. The protocols were customized slightly for the specific WIDA MODEL folders that were being studied.

Data from the cognitive lab included video and the notetakers' notes of student responses. After the pilot of the protocol and procedures, test developers slightly revised the possible codes for students' answers. The code "Student admittedly did not understand question or response options and guessed." was removed. The new codes "Student answered correctly based on key language." and "Student expressed unfamiliarity with or misinterpretation of theme/expected background knowledge." were added. After test developers coded students' responses to the items in Cognitive Lab 1, they met to discuss their codes and to reach a consensus about any item revisions.

Based on these analyses, test developers made suggestions for revising the Listening and Reading items. Some examples of their suggestions were paraphrasing technical language, clarifying confusing wording, making distractors more or less distracting, making graphics clearer, and creating a connection between the theme, passage, pictures, and items in a folder. Dorry Kenyon, Director of CAL's Language Testing Division, reviewed the suggested revisions and instructed test developers to revise certain items. After test developers made these revisions, they decided whether to further test the items through Cognitive Lab 2 or the field test.

2.7.1.3. Cognitive Lab 2

From September 21, 2010 through October 15, 2010, CAL employees and consultants conducted Cognitive Lab 2. This second cognitive lab included 25 students in Grades 6–8 from Paul Public Charter School in Washington, DC and Glasgow Middle School in Alexandria, VA and 54 students in Grades 9–12 from Columbia Heights Educational Complex (Bell High School) in Washington, DC, JEB Stuart High School in Falls Church, VA, and Liberty Middle School in Clifton, VA.

The main goals of the cognitive lab were to confirm that revisions made to Listening and Reading items after Cognitive Lab 1 were appropriate and that the items were performing at the intended proficiency levels. Prior to Cognitive Lab 2, interviewers and notetakers participated in a training session. They practiced interviewing, note-taking, and coding. Because the notes taken during the pilot test and Cognitive Lab 1 provided the most useful information, students were not video recorded in Cognitive Lab 2.

The interview and observation protocols in Cognitive Lab 2 were similar to those used in the pilot test and Cognitive Lab 1. Analyses used data from the notetakers' record of student responses. Test developers looked for trends in students' answers to items and reviewed students' feedback in the notes. The test developers, making generalizations based on the quantitative data and what the students said, created codes similar to those in the previous cognitive lab. The coders met to reach a consensus about the codes and on any possible revisions to items. Suggested revisions included paraphrasing sentences to remove difficult vocabulary, moving or removing parts of a picture, revising question options, and editing a passage so it related more clearly to the questions. Dorry Kenyon of CAL reviewed the suggested revisions and gave consent for test developers to make certain revisions.

2.7.2. Speaking

During development of the Speaking tests, employees at CAL conducted two rounds of pilot testing: 1) a cognitive lab with students in Grades 9–12 and 2) a cognitive lab with children of CAL staff members in Grades 6–8.

2.7.2.1. Speaking 9-12 SIL Cognitive Lab

On June 9, 2010, CAL employees conducted a cognitive lab for the 9-12 Speaking test. They tested and interviewed six high-proficiency ELLs who were in Grades 10 and 11 at Wilson High School in Washington, DC.

The goal of this cognitive lab was to test new Speaking tasks that were written at Proficiency Levels 4 and 5 on the SIL Standard to determine if they elicited language at Proficiency Levels 4 and 5 on the WIDA Speaking Rubric. These new tasks had been added to the end of a three-task folder, “Fine Arts Activities,” which had been retired from ACCESS. The previous three-task folder measured English language proficiency at only Proficiency Levels 1–3, so tasks needed to be added and piloted for Proficiency Levels 4–5. The Speaking test form for the cognitive lab also included a five-task folder, “Cars and Air Pollution,” which was a LoSS folder that had also been retired from ACCESS. This folder contained tasks at all five proficiency levels and had exhibited good psychometric statistics from operational testing, so the test developers could compare students’ performances on it during the cognitive lab with the same students’ performance on the other folder.

At the beginning of the cognitive lab, the interviewer administered the Speaking tasks to a student, and the student provided responses. In an observation protocol document, an observer took notes about the test administration, paying special attention to the student’s score for each task according to the WIDA Speaking Rubric (i.e., Meets, ?, or Approaches), the student’s behavior, and the test administrator’s behavior.

Then, the interviewer debriefed each student with a cognitive interview protocol that was based on the protocol that had been used for the ACCESS for ELLs pilot study. The observer recorded the student’s answer to each interview question in the observation protocol document. Specific questions from the interview protocol were as follows:

- Response
    - «Follow up on any specific aspects of the student’s response or behavior noted during the administration of the tasks.»
    - Did you like talking about [topic]?
    - We talked about [topic].Was it easy or hard to think of things to say?
    - You didn’t say very much about [topic]. Why not?

- Task
    - Have you talked about [topic] in school before?
    - Do you think other kids in your class would like talking about [topic]?
    - Do you think they would understand all the questions?
- Graphic
    - Do you like this picture? Why or why not?
    - What parts of the picture helped you answer the question?
    - Is there anything we should add or take away from the picture?
    - Was it easy or hard to understand what the picture shows?
- Additional issues

After the cognitive lab ended, test developers totaled the number of students scored as “Meets” and the number scored as “Approaches” for each Speaking task. The high-proficiency students easily met expectations for the first three tasks in the “Fine Arts Activities” folder, and most of those students met expectations for the two new Speaking tasks, indicating that these tasks were eliciting language at Proficiency Levels 4 and 5 as intended. According to the interview notes, students understood the tasks and graphics and were able to speak about the topics. In addition, students’ performances on this folder were similar to their performances on the folder “Cars and Air Pollution,” providing further evidence that “Fine Arts Activities” was measuring language at the appropriate levels.

In October 2010, test developers organized a cognitive lab for the 6-8 Speaking test. For this cognitive lab, called the “CAL Kids Cognitive Lab,” test developers recruited a few CAL staff parents and their children who were currently enrolled in Grades 6–8 in a Washington, DC-area school. The goal of this cognitive lab was to test new Speaking tasks that were written at Proficiency Levels 4 and 5 on the SIL Standard to determine if they elicited language at levels 4 and 5 on the WIDA Speaking Rubric. These new tasks had been added to the end of a three-task folder, “Book Club,” which had been retired from ACCESS. The previous three-task folder measured English language proficiency at only Proficiency Levels 1–3, so tasks needed to be added and piloted for Proficiency Levels 4–5. In addition, test developers were interested in seeing if passages were easy to read and understand, if questions were clear and elicited sufficient quality and quantity of language, if graphics were easy to see and relevant, and if tasks scaffolded to subsequent tasks. Data would be used to make improvements to the tasks before the field test was conducted.

Immediately prior to the cognitive lab, the test developers trained the parents to use an interview protocol and procedures. The parent then administered the cognitive lab to the child in their home. First, the parent described the overall process of conducting the cognitive lab. Then, the parent administered the Speaking folder to the child. Using the WIDA Speaking Rubric, the parent scored the students’ answers as “Meets,” “?,” “Approaches,” “No Response,” or “Not Administered.” The parent recorded the answers and scores in a document.

After the tasks had been administered, the parent conducted the cognitive interview. For each task, the parent followed the questions listed in the interview protocol and took detailed notes of the child’s answers to the questions. Specific questions from the interview protocol were as follows:

- Response
    - «Follow up on any specific aspects of the student’s response or behavior noted during the administration of the tasks.»
    - Did you like talking about [topic]?
    - We talked about [topic].Was it easy or hard to think of things to say?
    - You didn’t say very much about [topic]. Why not?
- Task
    - Have you talked about [topic] in school before?
    - Do you think other kids in your class would like talking about [topic]?
    - Do you think they would understand all the questions?
- Graphic
    - Do you like this picture? Why or why not?
    - What parts of the picture helped you answer the question?
    - Is there anything we should add or take away from the picture?
    - Was it easy or hard to understand what the picture shows?
- Additional issues
    - I noticed you paused at this question. Did you have a hard time thinking of something to say?
- Item-specific
    - «Specific_Questions»

The parent submitted the child’s answers and the interview notes to the test developers. Test developers computed how many total students scored “Meets” and how many scored “Approaches” for each Speaking task. The high-proficiency students were found to easily meet expectations for the first three tasks in the “Book Club” folder as well as for the two new Speaking tasks, suggesting that these tasks could elicit language at Proficiency Levels 4 and 5. According to the interview notes, students understood the tasks and graphics and were able to speak about the topics. Test developers discussed possible revisions to the Speaking tasks. Test developers had no large concerns about the test and therefore prepared it for the field test.

2.8. Finalizing the WIDA MODEL Field Test Forms

Pre-field test key checks were conducted by CAL employees and external consultants in mock test administrations on November 1, 2010. A final key check was conducted at CAL in June 2011 after the field test. A Project Coordinator oversaw the process, including fails and reconciliation steps.

Prior to publication, the WIDA MODEL test forms were proofed by test developers on soft copy and hard copy, by test developers acting as script readers and test takers in a mock test administration, and by an external professional editor. A CAL manager performed a final review of colored hard copies of all test materials, which were then approved by Dorry Kenyon.

Table 9 and Table 10 list the final folders that appear on WIDA MODEL for Grades 6–8 and 9–12, respectively. The tables also list the source of each folder, which WIDA ELP Standard it meets, the tier of the folder, and in which step (Step 1 or Step 2: Low, Mid, or High) it appears on the final test form. As mentioned earlier in this chapter, folders were generated from a variety of sources, including retired folders from ACCESS (see Chapter 2.1.), folders taken from the ACCESS field test (see Chapter 2.1), and newly created folders designed by item writers (see Chapter 2.2). Folders that appear in multiple placement levels are marked as “repeated.”

Table 9

List of Final MODEL Folders Grades 6–8

Domain and Folder Title	Source[7]	WIDA ELP Standard[8]	Folder Tier[9]	Step and Placement Level
Speaking				
Book Club	ACCESS 200 + CAL in-house	SIL	N/A	N/A
Amelia Earhart	ACCESS 103	LoSS	N/A	N/A
Listening				
The Traveler	ACCESS FT	LoLA	B+	Step 1
Poster Project	External item writer	SIL	A	Step 2: Low
Tara at the Art Museum	ACCESS 102	LoLA	A	Step 2: Low
Buying Books	ACCESS 200	LoMA	A	Step 2: Low
Growing Tomatoes	ACCESS 200	LoMA	B	Step 2: Mid
Exploring the Solar System	ACCESS 101	LoSC	B	Step 2: Mid
Railroads	ACCESS 200	LoSS	C	Step 2: Mid
The Hungry Coat (repeated)	CAL in-house	LoLA	C	Step 2: Mid
Buying Candy	ACCESS 101 revised	LoMA	C	Step 2: High
Science Tools	ACCESS 102	LoSC	C	Step 2: High
Renewable Energy	CAL in-house	LoSS	C	Step 2: High
The Hungry Coat (repeated)	CAL in-house	LoLA	C	Step 2: High
Hanging Scale	CAL in-house	LoMA/LoSC	C	Step2: Screener
Writing				
New School Club	ACCESS 103	IT	N/A	N/A
Mural Ideas	ACCESS 102 revised	IT	N/A	N/A
Reading				
Restaurant Review	ACCESS FT	LoLA	B+	Step 1
Cooking Eggs	External item writer	LoLA	A	Step 2: Low
Covering a Box	ACCESS 102	LoMA	A	Step 2: Low
Convection Currents	External item writer	LoSC	A	Step 2: Low
School Store	ACCESS FT	LoMA	B	Step 2: Mid
How Plants Make Their Food	ACCESS 101	LoSC	B	Step 2: Mid
The Industrial Revolution	External item writer	LoSS	B	Step 2: Mid
Winter Sun (repeated)	ACCESS 200	LoLA	C	Step 2: Mid
Book Sale	ACCESS 102	LoMA	C	Step 2: High
Pancakes	CAL in-house	LoSC	C	Step 2: High
Photography Firsts	External item writer	LoSS	C	Step 2: High
Winter Sun (repeated)	ACCESS 200	LoLA	C	Step 2: High
Chesapeake Bay	External item writer	LoMA/LoSC	C	Step 2: Screener

[7] Series 101 of ACCESS was administered operationally during the 2005–2006 academic year, Series 102 during 2006–2007, Series 103 during 2007–2008, Series 200 during 2008–2009, and Series 201 during 2009–2010.

[8] As mentioned previously, "SIL" is Social and Instructional Language, "LoLA" is Language of Language Arts, "LoMA" is Language of Mathematics, "LoSC" is Language of Science, and "LoSS" is Language of Social Studies. For Writing, "IT" integrates SIL, LoLA, and LoSS.

[9] The folder tier correlates with the proficiency level of items rather than to the placement level of the test (i.e., Low, Mid, and High). Tier A folders have items at Proficiency Levels 1–3. Tier B folders have items at Proficiency Levels 2–4. Tier C folders have items at Proficiency Levels 3–5. Tier B+ folders have items at Proficiency Levels 2–5.

Table 10

List of Final MODEL Folders Grades 9–12

		WIDA ELP	Folder	Step and Placement
Domain and Folder Title	Source	Standard	Tier	Level
Speaking				
Fine Arts Activities	ACCESS 103 + CAL in-house	SIL	N/A	N/A
Cars and Air Pollution	ACCESS 103	LoSS	N/A	N/A
Listening				
Group Behavior	CAL in-house	LoLA/LoSS	B+	Step 1
Sources of Information	ACCESS 101	SIL	A	Step 2: Low
One Day After School	External item writer	LoLA	A	Step 2: Low
Camilla’s Plant	CAL in-house	LoMA	A	Step 2: Low
Mr. Lee’s Store	ACCESS 101	LoMA	B	Step 2: Mid
Blue Crabs	External item writer	LoSC	B	Step 2: Mid
Political Alliances	ACCESS 200	LoSS	C	Step 2: Mid
Sea Story (repeated)	External item writer	LoLA	C	Step 2: Mid
Statistics	CAL in-house	LoMA	C	Step 2: High
Single-celled Organisms	CAL in-house	LoSC	C	Step 2: High
European Explorers	ACCESS 103	LoSS	C	Step 2: High
Sea Story (repeated)	External item writer	LoLA	C	Step 2: High
Balance	External item writer	LoMA/LoSC	C	Step 2: Screener
Writing				
Shirley Chisholm	ACCESS 200	IT	N/A	N/A
Babe Didrikson Zaharias	ACCESS 201	IT	N/A	N/A
Reading				
The Northern Sea	ACCESS 200	LoLA	B+	Step 1
Julia Child	CAL in-house	LoLA	A	Step 2: Low
Polygons	ACCESS FT	LoMA	A	Step 2: Low
Dietary Guidelines	ACCESS 101	LoSC	A	Step 2: Low
Angles in Four-Sided Shapes	ACCESS 102	LoMA	B	Step 2: Mid
Model Rockets	External item writer	LoSC	B	Step 2: Mid
Ancient Writing	ACCESS FT	LoSS	B	Step 2: Mid
Conservationist Joy Adamson (repeated)	ACCESS FT	LoLA	C	Step 2: Mid
Perspective	CAL in-house	LoMA	C	Step 2: High
Bacterial Growth	CAL in-house	LoSC	C	Step 2: High
Ancient Civilizations	ACCESS 101	LoSS	C	Step 2: High
Conservationist Joy Adamson (repeated)	ACCESS FT	LoLA	C	Step 2: High
The Electrical Circuit	External item writer	LoMA/LoSC	C	Step 2: Screener

2.9. *Selection of Folders for MODEL Screener*

In developing MODEL Screener for Grades 6–8 and 9–12, test developers commented that the Speaking sections in MODEL were already quite short. They also noticed that limiting the Writing sections in MODEL to only Part A in the Screener would not provide a sufficient measure of academic English language proficiency because Part A was not designed to elicit extended discourse. Therefore, the Speaking and Writing sections of MODEL Screener remained identical to those in MODEL. However, test developers noted that the more lengthy nature of the Listening and Reading sections in MODEL, which contained three or four folders for each placement level, allowed for them to be shortened in MODEL Screener.

In order to shorten the Listening and Reading sections for each grade-level cluster, test developers automatically included the Step 1 folder for each domain and then created one new Step 2 folder for each domain. Rather than using a shortened version of the Mid and High level folders from the full-length MODEL, the Step 2 folders in the MODEL Screener 6-8 and 9-12 test forms were specifically developed. Test developers determined that having a different folder for the Screener would extend the life of the test for students, allowing students to take the Screener at the beginning of the school year and then the full-length MODEL assessment at the end of the school year. Table 11 shows a list of the folder titles, sources, WIDA ELP Standards, folder tiers, steps, and placement levels that were included in the final version of MODEL Screener for Grades 6–8.

Table 11
List of Final MODEL Screener Folders for Grades 6–8

Domain and Folder Title	Source[10]	WIDA ELP Standard[11]	Folder Tier[12]	Step and Placement Level
		Speaking		
Book Club	ACCESS 200 + CAL in-house	SIL	N/A	N/A
Amelia Earhart	ACCESS 103	LoSS	N/A	N/A
		Listening		
The Traveler	ACCESS FT	LoLA	B+	Step 1
Hanging Scale	CAL in-house	LoMA/LoSC	C	Step2: Screener
		Writing		
New School Club	ACCESS 103	IT	N/A	N/A
Mural Ideas	ACCESS 102 revised	IT	N/A	N/A
		Reading		
Restaurant Review	ACCESS FT	LoLA	B+	Step 1
Chesapeake Bay	External item writer	LoMA/LoSC	C	Step 2: Screener

Table 12 shows a list of the folder titles, sources, WIDA ELP Standards, folder tiers, steps, and placement levels that were included in the final version of MODEL Screener for Grades 9–12.

[10] Series 101 of ACCESS was administered operationally during the 2005–2006 academic year, series 102 during 2006–2007, series 103 during 2007–2008, series 200 during 2008–2009, and series 201 during 2009–2010.

[11] As mentioned previously, SIL is Social and Instructional Language, LoLA is Language of Language Arts, LoMA is Language of Mathematics, LoSC is Language of Science, and LoSS is Language of Social Studies. For Writing, IT integrates SIL, LoLA, and LoSS.

[12] The folder tier correlates with the proficiency level of items rather than to the placement level of the test (i.e., Low, Mid, and High). Tier A folders have items at Proficiency Levels 1–3. Tier B folders have items at Proficiency Levels 2–4. Tier C folders have items at Proficiency Levels 3–5. Tier B+ folders have items at Proficiency Levels 2–5.

Table 12

List of Final MODEL Screener Folders for Grades 9–12

Domain and Folder Title	Source	WIDA ELP Standard	Folder Tier	Step and Placement Level
Speaking				
Fine Arts Activities	ACCESS 103	SIL	N/A	N/A
Cars and Air Pollution	ACCESS 103	LoSS	N/A	N/A
Listening				
Group Behavior	CAL in-house	LoLA/LoSS	B+	Step 1
Balance	External item writer	LoMA/LoSC	C	Step 2: Screener
Writing				
Shirley Chisholm	ACCESS 200	IT	N/A	N/A
Babe Didrikson Zaharias	ACCESS 201	IT	N/A	N/A
Reading				
The Northern Sea	ACCESS 200	LoLA	B+	Step 1
The Electrical Circuit	External item writer	LoMA/LoSC	C	Step 2: Screener

3. Field Test

3.1. Design of the Field Test

The field test for WIDA MODEL Grades 6-8 and 9-12 was conducted in the winter of 2010. The purpose of the field test was to collect data on items and tasks in order to examine their psychometric properties, to link WIDA MODEL field test scores to ACCESS operational scores for Listening and Reading, to link students' performances to WIDA ELP levels, and to analyze the validity and reliability of the tests. Only schools from the WIDA Consortium that were not already participating in the ACCESS field test took part in the WIDA MODEL field test, as ACCESS is administered only in member states and conducting two simultaneous field tests at a school was deemed too difficult. Test developers planned to administer WIDA MODEL to students a short period of time prior to the operational ACCESS administration.

To have sufficient data to conduct psychometric analyses, test developers aimed to assess approximately 300 students for each placement level of Low, Mid, and High in each grade-level cluster. Because of the variety of geographic locations of the schools and the number of students who needed to be assessed, CAL staff decided upon a cost- and time-effective plan to train a group of field test administrators (FTAs), who were local to the participating schools and would administer WIDA MODEL to students. This plan is described in more detail in the subsequent sections of this chapter.

3.2. Participation Data

WIDA MODEL was field tested in four WIDA states—Illinois, Kentucky, Maine, and New Mexico—from November through December 2010. Table 13 lists the schools that participated in the field test. Twenty-four schools from 10 school districts participated.

Table 13

Schools that Participated in the 2010 WIDA MODEL Field Test

Dates	District	School
November 29 – December 3, 2010	Aurora West USD 112, IL	Washington Middle School
November 29 – December 10, 2010	Aurora West USD 112, IL	West Aurora High School
November 29 – December 10, 2010	Bowling Green Independent SD, KY	Bowling Green High School
December 6–17, 2010	Bowling Green Independent SD, KY	Bowling Green Junior High School
December 6–10, 2010	Cicero SD99, IL	Unity Junior High School
November 15–16, 2010	Elgin SD U46, IL	Elgin High School
November 17–18, 2010	Elgin SD U46, IL	Ellis Middle School
November 29–30, 2010	Elgin SD U46, IL	Larkin High School
November 17–18, 2010	Elgin SD U46, IL	Larsen Middle School
November 29–30, 2010	Elgin SD U46, IL	Streamwood High School
November 29 – December 17, 2010	Gallup-McKinley SD, NM	Gallup High School
November 29 – December 10, 2010	Gallup-McKinley SD, NM	JFK Middle School
November 29 – December 17, 2010	Gallup-McKinley SD, NM	Tohatchi High School
November 15–19, 2010	Glen Ellyn SD 87, IL	Hadley MS
November 15 – December 7, 2010	Plainfield SD 202, IL	Aux Sable Middle School
November 15 – December 3, 2010	Plainfield SD 202, IL	Indian Train Middle School
November 15 – December 1, 2010	Plainfield SD 202, IL	Plainfield Central High School
November 15–19, 2010	Portland SD, ME	Lincoln Middle School
November 15–19, 2010	Portland SD, ME	Lyman Moore Middle School
November 29 – December 10, 2010	Warren County Central SD, KY	Greenwood High School
November 29 – December 10, 2010	Warren County Central SD, KY	Warren Central High School
December 3–10, 2010	Wheeling CCSD 21, IL	Cooper Middle School
November 15–19, 2010	Wheeling CCSD 21, IL	Oliver Holmes Middle School
November 29 – December 13, 2010	Wheeling CCSD 21, IL	Jack London Middle School

To recruit the FTAs, CAL advertised on Craigslist, listservs, and newspapers in metropolitan areas close to the schools. Consultants were hired based on interviews with CAL test developers. Qualified applicants had experience working with ELL and/or secondary students, had strong interpersonal skills, had native or near-native English proficiency, had access to a reliable car, provided proof of a background check, and were available for training and working during certain days and hours.

Prior to the start of the field test, selected applicants were sent test administration training materials, including test administration manuals and Microsoft PowerPoint presentations, which they were required to read independently. Later, CAL held a full-day, in-person training for the test administrators at a location central to them. The training consisted of signing nondisclosure agreements, reviewing the test materials and format, practicing the administration and scoring of each domain section on MODEL and MODEL Screener, and learning general field test procedures.

As seen in Table 14, 51 FTAs from Illinois, Kentucky, Maine, and New Mexico and five employees from CAL were responsible for administering the WIDA MODEL field test. In New Mexico, due to the large number of potential students and the difficulty of finding and training qualified test administrators, CAL sent some of its own staff to serve as test administrators. CAL also sent one FTA from Maine to administer WIDA MODEL in New Mexico.

Table 14
Field Test Administrators

State	Field Test Administrators
IL	Ayesha Ahmed, Barbara Thomases, Bianca Greenwald, Brandon Oswald, Elana Jacobs, Georgia Deep, Gina Orazi, Jasper Phillips, Jeanne Rothlisberger, Jennifer Gebberhardt, Jennifer Ruzich, Judith Ball, Kathleen Corso, Kathleen Gomez, Kelly Whitehead, Kristin Huzar, Linda Haseman, Linda Rosenquist, Lindsay Curry, Marcy Goodman, Mary Crambes, Maureen O'Brien, Michael Soto, Noreen Haque, Rachel Benedict, Sabrina Kaiser, Sarah Linsey, Sue Varava, Susan Stieber, and Suzanne Edwards
KY	Amanda (Kate) Scott, Anna Michalak, Charlotte May, Dierdre Rieppel, Gretchen Collins, Inga Wolff, Lynne Croxton, Margaret Conrad, Michael Birdsall, Sandy Mefford, Sheila Duncan, and Thamara Rhodes
ME	Amanda Wogaman, Amy Temple, David Spear, Linda Hoffman, and Marlies Reppenhagen
NM	Amy Temple (ME), Anna Todorova (CAL), Carolene Whitman, Deepak Ebenezer (CAL), Jacqueline Lopez (CAL), Rose Wyaco, Sheryl Smith, Stephanie Gibson (CAL), and Tatyana Vdovina (CAL)

The WIDA MODEL series was field tested on a total of 1,256 public school students in Grades 6–12. Table 15 shows the demographic characteristics of the students by grade-level cluster. To obtain this information, the students who took WIDA MODEL were matched to the students who took the operational ACCESS Series 202 (2010–2011 academic year) test. The students who took only WIDA MODEL and not ACCESS are counted as "Missing" in the table. For grade-level cluster 6-8, the sample consisted of slightly more female than male students, and the majority of the students were of Hispanic ethnicity and from the state of Illinois. For grade-level cluster 9-12, the sample consisted of slightly more male than female students and showed more diversity with respect to race/ethnicity and state. The largest groups of students were Hispanic, Non-Hispanic White, and Non-Hispanic American Indian. Students were drawn from New Mexico, Kentucky, or Illinois in fairly equal numbers.

Table 15
Demographics for the Field Test Students by Grade-level Cluster

	Grade-level Cluster 6-8		Grade-level Cluster 9-12	
	N	P	N	P
Sex				
Female	364	49.9%	221	42.0%
Male	301	41.2%	268	51.0%
Missing	65	8.9%	37	7.0%
Race/Ethnicity				
Non-Hispanic Asian	44	6.0%	125	23.8%
Non-Hispanic Pacific Islander	1	0.1%	2	0.4%
Non-Hispanic Black	25	3.4%	17	3.2%
Hispanic (Of Any Race)	492	67.4%	136	25.9%
Non-Hispanic American Indian	50	6.8%	168	31.9%
Non-Hispanic White	51	7.0%	38	7.2%
Missing	67	9.2%	40	7.6%
State				
Illinois	608	83.3%	156	29.7%
Kentucky	28	3.8%	170	32.3%
Maine	35	4.8%	0	0.0%
New Mexico	58	7.9%	200	38.0%
Missing	1	0.1%	0	0.0%

3.3. Administration of the Field Test

Prior to the start of the field test, CAL staff members finalized schedules with school liaisons. Schools submitted the names and some demographic information of students who would participate, and MetriTech printed labels for each student's test booklet in order to be able to match WIDA MODEL data to ACCESS data.

CAL prepared all test materials for the field test administration. One test booklet contained the Speaking and Listening prompts, and another test booklet contained the Reading passages. The Student Response Booklets contained the sections for the test administrators to record the students' answers, the writing prompts, and space for the students to fill in their Listening and Reading answers. A test administration manual was prepared for all test administrators. Scripts containing the passages to be read to the students for the Speaking and Listening sections and the instructions for the Writing and Reading sections were also included. Materials were personally delivered to schools by CAL test developers. CAL test developers would observe test administrations on that first day in order to ensure that procedures went smoothly.

The field test and operational test administration for WIDA MODEL were largely based on the test administration for ACCESS. For example, general room setup and testing procedures included arranging the desks in the testing room so students could see and hear the test administrator, as well as ensuring that students had sharpened pencils, that a Do Not Disturb sign was placed on the door, that a watch or clock was available to pace the test, that test materials were distributed to the correct student, that test materials were kept secure, and that the test administrator's script was followed exactly. As would occur in the operational test, the WIDA MODEL field test was administered by the FTAs in the following sequence: Speaking, Listening, Writing, and Reading. (See Chapter 1.4 for details about the administration of the test.)

CAL recommended that students be tested in one session for Speaking and Listening and then a second session for Writing and Reading. FTAs were strongly advised to give the Writing and Reading domains to students in a small group, as allowed by the Test Administration Manual. Because this was a field test and participation was voluntary, students who missed one of the two testing sessions were not required to make up that part of the assessment. After the FTAs had administered the tests to students, they copied students' answers to scannable score sheets and returned all materials to CAL.

3.4. Scoring Procedures

The following sections of this report summarize the procedures for scoring students' responses during the field test administration. These procedures are similar to the scoring procedures that later became operational.

3.4.1. Scoring the Speaking Section

After each task during the administration of the Speaking test, the field test administrator made a qualitative judgment about the student's performance by assigning one of the following possible ratings:

- Meets,
- ? (question mark), or
- Approaches.

"Meets" indicates that the student's response meets or exceeds all task level expectations in quantity and quality. "Approaches" means that the student approaches task level expectations but falls short in quantity and/or quality, gives no response, or gives a response in a language other than English. A question mark means that the test administrator is unsure if the student's response is "Meets" or "Approaches." In such cases, the test administrator moves on to the next task and then returns to score the response as "Meets" or "Approaches" based on the student's subsequent response.

The Speaking tasks were developed to allow students to give a performance at each proficiency level as defined in the WIDA Consortium's Speaking Rubric. A student's response was not judged on whether the content was right or wrong but rather on whether it met the language proficiency level expectations for each task on three criteria—Linguistic Complexity, Vocabulary Usage, and Language Control. For example, if a student gave a response that did not address the content of the question, but that response still met the proficiency-level expectations of the task, it was scored as "Meets." The total Speaking raw score for a student was the sum of every response that was scored as "Meets."

3.4.2. Scoring the Listening Section

During administration of the Listening section, as each student pointed to or said aloud his or her answers to multiple-choice items, the test administrator recorded the answers in the Student Response Booklet. The test administrator scored the Listening Step 1 items as correct or incorrect based on answer keys during test administration because these scores, in conjunction with the Speaking scores, determined placement of Low, Mid, or High for Listening Step 2. Then, the test administrator administered the Step 2 placement items and recorded the student's responses. After the student had left the testing area, the test administrator used the answer keys to score the Step 2 placement items as correct or incorrect. The test administrator summed the total number of correct answers to calculate the raw score.

3.4.3. Scoring the Writing Section

During the test administration, the test administrator gave students a Student Response Booklet that contained either Writing Task 1 or Writing Task 2. The tasks are about different topics and require different types of writing, but either task could be administered to a student. Each task consists of two parts: Part A is a prompt that allows students to use single words, phrases, or simple sentences to describe what they see in a picture, and Part B is a prompt that allows students to construct an extended narrative response. A student attempted Part B only if he or she met performance expectations in Part A.

In the field test, after the student completed the Writing test and Reading Step 1, the test administrator used the scoring criteria to quickly evaluate the student's writing by assigning a Quick Score of Low, Mid, or High. A response received a Quick Score of Low if the student completed only Part A or if he or she produced only single words or copied text on Part B. A response received a Quick Score of High if the student wrote a well-organized composition that used a variety of sentence lengths, contained specific and technical vocabulary, and was easily comprehended. A response that exceeded the criteria for Low but did not meet the criteria for High was scored as Mid. Along with the number of correct responses in Reading Step 1 (see

Chapter 3.4.4), this Quick Score determined whether a student proceeded to Low, Mid, or High in Reading Step 2.

After the test administrators concluded their work on the field test and returned all materials to CAL, students' writing samples were scored by CAL staff and consultants according to the WIDA Consortium's Writing Rubric (see Chapter 4.2.1).

3.4.4. Scoring the Reading Section

The Reading section consists of a series of reading passages followed by multiple-choice questions. Each student filled in his or her own answers in the Student Response Booklet. When each student finished responding to the four items in Reading Step 1, the test administrator checked the responses against the answer key. In addition, the test administrator skimmed the Writing Section to obtain the Writing Quick Score of High, Mid, or Low. The Reading Step 1 score, together with the Writing Quick Score, determined the appropriate placement of Low, Mid, or High for Reading Step 2. The student then marked his or her answers for Reading Step 2 in the Student Response Booklet. After the student had left the testing area, the test administrator marked each item in Reading Step 2 as correct or incorrect using the answer keys. Then he or she recorded the total number of correct answers (i.e., total raw score), as well as the student's placement, in the Student Response Booklet.

3.5. Data Cleaning

Before the field test data for WIDA MODEL were finalized, multiple stages of data cleaning, processing, and quality checks were conducted. Upon receipt of testing materials, test developers read the test administrator report forms to see if there were any problems with the testing sessions, and, if so, flagged the appropriate sections in the Student Response Booklets and determined next steps for any data cleaning. Next, test developers checked that the student responses that the test administrators had copied over to the scannable score sheets matched the students' responses in the test booklets. Incorrect bubbles, sections that were accidentally left blank, and stray marks were corrected.

CAL research assistants scanned the scannable forms with Gravic® Remark Office OMR® software and a scanner to electronically collect data. Each scannable form was scanned twice in case the scanner malfunctioned. Data were exported from Remark to Microsoft Excel, where data were cleaned by comparing the data from each student's two scannings. Discrepancies between each student's scannings were manually corrected as necessary. After the field test data were finalized, students' original responses for all domains except for Writing were converted into numeric values for psychometric analysis. For Speaking tasks, "Meets" was coded as "1" and "Approaches" as "0". For Listening and Reading items, correct answers were coded as "1" and incorrect answers were coded as "0."

Students' scored responses on WIDA MODEL items and tasks were cleaned in two phases. In Phase 1, students with missing responses on MODEL or MODEL Screener were removed. This dataset was used to conduct the demographic analyses of the WIDA MODEL Listening and Reading domains.

Table 16 shows the breakdown of students in the Phase 1 data cleaning by grade-level cluster, domain, and placement level of test form for Listening and Reading on each placement level of MODEL (i.e., Low, Mid, High) and MODEL Screener.

Table 16 also shows the number of students whose responses were excluded from analyses because of incomplete data for MODEL or MODEL Screener. The students with "Missing MODEL Responses" or "Missing Screener Responses" had, by test design or test administration error, completed only Step 1 and did not complete either the placement step or the Screener. Only students with "Complete MODEL scores" were included in the demographic analyses.

Table 16
Results of Phase 1 Data Cleaning for Listening and Reading

			Number of Students per Level of Test Form				Number of Students with Incomplete MODEL Scores		Number of Students per Level of Test Form with Complete MODEL Scores			
Grade-level Cluster	Domain	Total Number of Students	Low	Mid	High	Screener	Missing MODEL Responses	Missing Screener Responses	Low	Mid	High	Screener
6–8	Listening	727	72	418	129	108	1(mid)	2	72	417	129	106
6–8	Reading	722	66	518	32	106	1(mid)	4	66	517	32	102
9–12	Listening	521	128	233	59	101	0	8	128	233	59	93
9–12	Reading	510	117	253	40	100	0	23	117	253	40	77

Table 17 shows the final count of Field Test students included in the analyses for Listening and Reading after Phase 2 of data cleaning. The table shows the students with complete Listening and Reading data for MODEL and MODEL Screener, and the students with complete data for both MODEL and ACCESS after outliers were removed. The students who took only WIDA MODEL and not ACCESS were counted as "Missing ACCESS Scores" in the table. (Details of the outlier analysis are described in Chapter 4.1.2.) Only students included in the dataset after both Phase 1 and Phase 2 of data cleaning were completed were used to conduct the Rasch analysis for the WIDA MODEL Listening and Reading domains.

Table 17
Results of Phase 2 Data Cleaning for Listening and Reading

		Number of Students per Level of Test Form with Complete MODEL Scores				Number of Students Excluded from Rasch Analysis		Number of Students per Level of Test Form Included in Rasch Analysis			
Grade-level Cluster	Domain	Low	Mid	High	Screener	Missing ACCESS Scores	Outliers	Low	Mid	High	Screener
6–8	Listening	72	417	129	106	139	32	53	330	93	77
6–8	Reading	66	517	32	102	137	43	53	390	20	74
9–12	Listening	128	233	59	93	47	21	113	203	46	83
9–12	Reading	117	253	40	77	42	17	103	225	32	68

The number of students for Speaking and Writing is shown in Table 18, as these sections of the test do not have placement levels and are the same for MODEL and MODEL Screener. Only students with complete data for both WIDA MODEL and ACCESS were used to conduct the Rasch analyses for the WIDA MODEL Speaking and Writing domains.

Table 18
Number of Field Test Students for Speaking and Writing

Grade-level Cluster	Domain	Total Number of Students	Number of Students Excluded from Rasch Analysis Due to Missing ACCESS Scores	Number of Students Included in Rasch Analysis
6–8	Speaking	728	139	589
6–8	Writing	723	N/A	723
9–12	Speaking	526	0	526
9–12	Writing	515	N/A	515

4. Field Test Results

This chapter presents the results of analyses conducted on data collected during the WIDA MODEL Field Test for Grades 6-8 and 9-12. For the Speaking, Listening, and Reading sections, Rasch analyses were used to examine how well the tasks and items function. For the Writing section, many-facet Rasch analyses were used to analyze the student responses.

4.1. Results for MODEL Speaking, Listening, and Reading

4.1.1. Descriptive Statistics for Speaking, Listening, and Reading

For the Speaking section of WIDA MODEL, raw scores range from 0–10. For the Listening and Reading sections, raw scores can range from 0–16, with the maximum possible score dependent upon the student placement level (13 for students in the Low placement level; 16 for students in the Mid or High placement levels). For quality assurance, researchers at CAL recomputed the field test administrators' total raw scores for each student. Descriptive statistics for the Speaking, Listening, and Reading sections for grade-level clusters 6–8 and 9–12 are presented in Table 19 and Table 20, respectively. Descriptive statistics are based on students with MODEL scores.

Table 19

Descriptive Statistics for MODEL Grades 6–8 by Step and Placement Level

Domain	Step and Placement Level	No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking	-	10	728	0	10	7.06	2.22
Listening	Step 1	4	727	0	4	2.45	1.06
	Step 2: Low	9	72	1	8	4.47	1.80
	Step 2: Mid	12	417	1	12	7.23	1.97
	Step 2: High	12	129	2	11	7.71	1.73
	Step 1 and Step 2: Low	13	72	1	10	5.78	2.25
	Step 1 and Step 2: Mid	16	417	2	16	9.56	2.39
	Step 1 and Step 2: High	16	129	6	15	11.18	1.84
Reading	Step 1	4	722	0	4	1.80	1.02
	Step 2: Low	9	66	1	9	4.86	1.82
	Step 2: Mid	12	517	1	12	6.20	2.23
	Step 2: High	12	32	1	12	6.81	2.58
	Step 1 and Step 2: Low	13	66	2	11	5.83	2.09
	Step 1 and Step 2: Mid	16	517	1	16	8.04	2.61
	Step 1 and Step 2: High	16	32	4	15	9.88	2.67

Table 20

Descriptive Statistics for MODEL Grades 9-12 by Step and Placement Level

Domain	Step and Placement Level	No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking	-	10	526	0	10	5.65	2.65
Listening	Step 1	4	526	0	4	2.04	1.09
	Step 2: Low	9	128	0	9	4.01	2.03
	Step 2: Mid	12	233	1	11	5.13	1.94
	Step 2: High	12	59	2	10	5.98	1.83
	Step 1 and Step 2: Low	13	128	1	11	5.22	2.28
	Step 1 and Step 2: Mid	16	233	2	14	7.28	2.19
	Step 1 and Step 2: High	16	59	5	14	9.31	1.87
Reading	Step 1	4	514	0	4	1.98	1.14
	Step 2: Low	9	117	0	9	4.62	1.76
	Step 2: Mid	12	253	0	11	5.34	2.29
	Step 2: High	12	40	2	11	5.45	1.83
	Step 1 and Step 2: Low	13	117	1	11	5.78	1.93
	Step 1 and Step 2: Mid	16	253	1	15	7.55	2.73
	Step 1 and Step 2: High	16	40	6	14	8.88	1.91

4.1.2. Outlier Analysis for Listening and Reading

One of the purposes of the WIDA MODEL field test is to link WIDA MODEL Listening and Reading scores to ACCESS Listening and Reading scores through the concurrent calibration method (see Chapter 5.1) so that performances on the WIDA MODEL Listening and Reading tests can be interpreted in terms of the WIDA ELP levels used for ACCESS. Since the goal of the concurrent calibration is to obtain the best estimates of WIDA MODEL parameters as if WIDA MODEL and ACCESS were administered at the same time and in the same testing conditions, students that performed very differently on WIDA MODEL and ACCESS were removed from the Rasch analysis and the linking analysis as their data would distort the linking results. There are several possible external reasons why some students may perform very differently on these two assessments. For example, students may not have tried their best when taking either WIDA MODEL or ACCESS. Or perhaps there are some subtle differences between the two testing conditions that were not under the researchers' control. For example, WIDA MODEL field test administrators made sure that all students attempted all the items while the ACCESS test administrators did not. Thus, students could potentially get very high scores on WIDA MODEL but very low scores on ACCESS simply because they did not attempt all of the ACCESS items.

The following procedure was used to identify and remove outliers from the Rasch analysis and linking analysis for Listening and Reading. First, an initial concurrent calibration was conducted to put WIDA MODEL on the ACCESS scale; then, student's WIDA MODEL and ACCESS logits scores were computed and compared. For each WIDA MODEL placement level, the mean and standard deviation of the students' difference in logits scores were computed. Then, students whose logits scores on MODEL and ACCESS differed by more than two standard deviations were identified and removed from the final datasets that were subsequently used for calibration and linking. Approximately 3-6% of student records were identified as outliers and were removed from the final dataset used for the Listening and Reading Rasch and linking analyses that appear in the remainder of this report. (Table 17 in Chapter 3.5 presents the number of students identified as outliers due to unexpectedly discrepant performances between the two administrations for Listening and Reading.)

4.1.3. Rasch Analyses for MODEL Speaking, Listening, and Reading

The dichotomous Rasch model operationalized in the Winsteps software program (software Version No. 3.70.0.5, Linacre, 2011) was used to analyze the test items for Speaking, Listening, and Reading. For all three domains, items were analyzed or calibrated in order to place items in a given grade-level cluster on the same scale. For the Reading and Listening domains, Step 1 and Step 2 items were estimated together in one Winsteps run.

Mathematically, the dichotomous Rasch model may be presented as

$$\log(\frac{P_{ni1}}{P_{ni0}}) = B_n - D_i$$

where

P_{ni1} = probability of a correct response by person *n* on item *i*
P_{ni0} = probability of an incorrect response by person *n* on item *i*
B_n = ability of person *n*
D_i = difficulty of item *i*

The Rasch model estimates the probability that a student will answer an item correctly given the difficulty of the item and the ability of the student. When the probability of a person getting a correct answer equals the probability of a person getting an incorrect answer (i.e., a 50 percent probability of getting it right and a 50 percent probability of getting it wrong), P_{ni1}/P_{ni0} is equal to 1. The log of 1 is 0. This is the point at which a person's ability equals the difficulty of an item. For example, if a person whose ability is 1.56 on the Rasch logit scale encounters an item whose difficulty is 1.56 on the Rasch logit scale, he or she would have a 50 percent probability of answering that item correctly. A logit is the unit of measurement used by Rasch for calibrating items and measuring persons.

Rasch models are confirmatory and assume a strong theoretical grounding for item development. Thus, measures that fit the measurement model may be considered, psychometrically speaking, to be very strong measures. Various Rasch item statistics for WIDA MODEL were computed and analyzed to examine whether items are considered strong measures of English language proficiency.

In this chapter of the report, the first column of each table shows the Item Name. Each part of each item name provides specific information about the item. For example, for the Speaking test for Grades 6–8, the name of the first item, “MODEL_S68_SI_p1_BookClub_A_T1,” represents the following: “MODEL” for the assessment, “S68” for the Speaking domain and the 6-8 grade level cluster, “SI” for the WIDA ELP Standard SIL, “p1” for the “Entering” proficiency level , “BookClub” for the folder title, “A” for the first folder or part, and “T1” for the first task.

The second column shows the Score, which is the number of Students who answered the item correctly.

The third column shows the Count, the total number of Students in the analysis for that item. This count of Students varies from folder to folder for Listening and Reading because different students took different folders depending on their placement level.

The fourth column shows the P-value of the item, which is the percentage of Students who answered the item correctly. The p-value was computed by dividing the Score by the Count. A p-value of 0.20 or less indicates a relatively difficult question, and a p-value of 0.80 or more indicates a relatively easy question.

The fifth column shows the Measure, the Rasch logit measure of the item. The Rasch measure for items is the item difficulty. A large, positive measure indicates a relatively difficult item, and a large, negative measure indicates a relatively easy item. These measures represent the final estimates for each item after anchoring them to their values based on common items. (See Chapter 5 for more details).

The sixth, IN.MSQ, and seventh, OUT.MSQ, columns are the infit and outfit mean square statistics. Infit and outfit statistics indicate any consistently unusual performance in relation to the item’s difficulty measure. They measure the degree to which students’ responses to items deviate from expected responses. Both statistics have an expected value of 1.0. The following criteria used to evaluate the infit and outfit mean square statistics should be regarded as relative as opposed to absolute criteria, as both statistics are affected by factors other than the quality of the measurement that the item produces: Items with infit and outfit mean square statistics between 0.5 and 1.5 are considered “productive for measurement” (Linacre, 2002). Values between 1.5 and 2.0 are “unproductive for construction of measurement, but not degrading.” Values greater than 2.0 might “distort or degrade the measurement system.” Values below 0.5 are “less productive for measurement, but not degrading.” Items (or students) with a higher-than-desirable fit statistic are referred to as misfitting, while items (or students) with a lower-than-desirable fit statistic are referred to as overfitting. Infit statistics are weighted, meaning they have been adjusted for outliers and are therefore less susceptible to inflation from inconsistent or unexpected responses than outfit statistics. Infit can be skewed if students within range of the targeted proficiency level do not perform as expected. Outfit is not weighted and therefore is very sensitive to outliers. Outfit can be skewed if students with extreme (i.e., high-level or low-level) proficiency do not perform as expected. High infit is a bigger threat to validity, but is more difficult to diagnose than high outfit (Linacre, 2002).

4.1.3.1. Rasch Analyses for the 6–8 Grade-level Cluster

Results of the Rasch item analysis for Grades 6–8 Speaking are reported in Table 21. The p-value and measure columns show that within the two folders—Part A and Part B—of the assessment: a) each consecutive Speaking task was more difficult than the one before it, as entailed by the test design; b) more students answered tasks correctly in the first folder versus the second folder; and c) consistent with the *a priori* proficiency levels used during task development, fewer and fewer students responded correctly to test tasks as the proficiency level increased. Such patterns are consistent with the adaptive test design.

Table 21
Rasch Item Analysis: Grades 6–8 Speaking

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_S68_SI_p1_BookClub_A_T1	586	589	0.99	-6.30	1.00	1.00
2.MODEL_S68_SI_p2_BookClub_A_T2	571	589	0.97	-3.33	0.57	0.04
3.MODEL_S68_SI_p3_BookClub_A_T3	513	589	0.87	-0.47	0.76	0.33
4.MODEL_S68_SI_p4_BookClub_A_T4	339	589	0.58	2.31	1.04	3.30
5.MODEL_S68_SI_p5_BookClub_A_T5	175	589	0.30	4.30	1.13	9.90
6.MODEL_S68_LS_p1_AmeliaEarhart_B_T1	578	589	0.98	-4.21	1.48	2.46
7.MODEL_S68_LS_p2_AmeliaEarhart_B_T2	550	589	0.93	-1.83	1.05	1.85
8.MODEL_S68_LS_p3_AmeliaEarhart_B_T3	441	589	0.75	0.97	0.76	0.49
9.MODEL_S68_LS_p4_AmeliaEarhart_B_T4	280	589	0.48	2.99	0.85	0.41
10.MODEL_S68_LS_p5_AmeliaEarhart_B_T5	133	589	0.23	4.96	0.97	0.33

Table 22 summarizes the infit and outfit findings. These infit and outfit mean square statistics are indicators of how well the data fit the Rasch measurement model. All ten tasks have infit mean square statistics that are between 0.5 and 1.5 and are productive for measurement (Linacre, 2002) of students’ speaking proficiency. For the outfit mean square statistics, one task has a value between 0.5 and 1.5, five tasks are less than 0.5, one task is between 1.5 and 2.0, and three tasks are greater than 2.0. For the three tasks with high outfit, the infit statistic falls within the productive range. Rasch users "routinely pay more attention to infit scores than outfit scores" (Bond & Fox, 2001, p. 43), as the infit statistics have been adjusted for outliers and are therefore less susceptible to inflation from inconsistent or unexpected responses, particularly for very easy or very difficult tasks (Linacre, 2002).

Table 22
Distribution of Mean-Square Fit Statistics: Grades 6–8 Speaking

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0	N = 0	N = 3
"distorting or degrading measurement"	% = 0%	% = 30%
> 1.5–2.0	N = 0	N = 1
"unproductive but not degrading"	% = 0%	% = 10%
0.5–1.5	N = 10	N = 1
"productive for measurement"	% = 100%	% = 10%
< 0.5	N = 0	N = 5
"less productive but not degrading"	% = 0%	% = 50%
	N = 10	N = 10
Total	% = 100%	% = 100%

Table 23 presents the results of the Rasch analyses on the 38 Listening items for grade-level cluster 6-8, and Table 24 summarizes the findings. The first four items are from Step 1; the latter items are for the Low, Mid, and High placement levels of Step 2 and for the Screener. The items with the label "COM" are common to both the Mid and High level forms. Infit statistics for 37 items are within the range that is considered productive for measurement, and 1 item was identified as unproductive but not degrading for measurement. According to the outfit statistics, 35 items are considered productive for measurement of students' listening proficiency, 2 items were identified as being unproductive but not degrading for measurement, and 1 item was identified as potentially distorting or degrading measurement. Further examination of these three items revealed unusual response patterns. Some of the students who incorrectly answered these three items got most other items in their placement level correct.

Table 23

Rasch Item Analysis: Grades 6–8 Listening

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_L68BC_LA_TheTraveler_p2_Step_1	433	558	0.78	0.42	0.99	1.00
2.MODEL_L68BC_LA_TheTraveler_p3_Step_1	269	558	0.48	1.91	1.03	1.01
3.MODEL_L68BC_LA_TheTraveler_p4_Step_1	279	558	0.50	1.82	1.02	1.01
4.MODEL_L68BC_LA_TheTraveler_p5_Step_1	376	558	0.67	1.00	0.99	0.95
5.MODEL_L68A_SI_PosterProject_p1_Low	40	53	0.75	-0.61	0.82	0.82
6.MODEL_L68A_SI_PosterProject_p2_Low	43	53	0.81	-0.99	0.82	0.70
7.MODEL_L68A_SI_PosterProject_p3_Low	23	53	0.43	1.05	1.30	1.69
8.MODEL_L68A_LA_Tara_p1_Low	28	53	0.53	0.60	1.10	1.08
9.MODEL_L68A_LA_Tara_p2_Low	31	53	0.58	0.32	0.96	0.88
10.MODEL_L68A_LA_Tara_p3_Low	33	53	0.62	0.13	0.78	0.69
11.MODEL_L68A_MA_BuyingBooks_p1_Low	5	53	0.09	3.32	0.99	2.25
12.MODEL_L68A_MA_BuyingBooks_p2_Low	32	53	0.60	0.23	0.85	0.81
13.MODEL_L68A_MA_BuyingBooks_p3_Low	7	53	0.13	2.91	1.01	0.80
14.MODEL_L68B_MA_GrowingTomatoes_p2_Mid	308	330	0.93	-1.00	0.97	0.82
15.MODEL_L68B_MA_GrowingTomatoes_p3_Mid	224	330	0.68	1.00	0.96	0.98
16.MODEL_L68B_MA_GrowingTomatoes_p4_Mid	96	330	0.29	2.79	1.03	1.13
17.MODEL_L68B_SC_SolarSystem_p2_Mid	223	330	0.68	1.02	0.94	0.93
18.MODEL_L68B_SC_SolarSystem_p3_Mid	241	330	0.73	0.74	1.02	0.96
19.MODEL_L68B_SC_SolarSystem_p4_Mid	165	330	0.50	1.82	1.00	0.99
20.MODEL_L68C_SS_Railroads_p3_Mid	183	330	0.55	1.58	0.96	0.94
21.MODEL_L68C_SS_Railroads_p4_Mid	116	330	0.35	2.49	1.10	1.07
22.MODEL_L68C_SS_Railroads_p5_Mid	163	330	0.49	1.85	0.95	0.94
23.MODEL_L68C_LA_HungryCoat_p3_COM	376	423	0.89	-0.32	1.03	1.05
24.MODEL_L68C_LA_HungryCoat_p4_COM	208	423	0.49	1.97	0.98	0.99
25.MODEL_L68C_LA_HungryCoat_p5_COM	252	423	0.60	1.51	1.03	1.08
26.MODEL_L68C_MA_BuyingCandy_p3_High	3	93	0.03	5.91	0.92	0.84
27.MODEL_L68C_MA_BuyingCandy_p4_High	84	93	0.90	-0.05	0.97	0.95
28.MODEL_L68C_MA_BuyingCandy_p5_High	18	93	0.19	3.86	1.25	1.59
29.MODEL_L68C_SC_ScienceTools_p3_High	3	93	0.03	5.91	1.11	1.11
30.MODEL_L68C_SC_ScienceTools_p4_High	35	93	0.38	2.87	1.01	0.80
31.MODEL_L68C_SC_ScienceTools_p5_High	27	93	0.29	3.29	0.96	0.93
32.MODEL_L68C_SS_RenewableEnergy_p3_High	90	93	0.97	-1.24	0.99	0.95
33.MODEL_L68C_SS_RenewableEnergy_p4_High	59	93	0.63	1.73	0.86	0.80
34.MODEL_L68C_SS_RenewableEnergy_p5_High	72	93	0.77	1.00	1.02	1.01
35.MODEL_L68C_MS_HangingScale_p3_Screener	50	77	0.65	1.19	0.67	0.85
36.MODEL_L68C_MS_HangingScale_p4_Screener	7	77	0.09	4.19	0.48	1.41
37.MODEL_L68C_MS_HangingScale_p5_Screener	21	77	0.27	3.00	1.12	0.75
38.MODEL_L68C_MS_HangingScale_p5_Screener	21	77	0.27	3.07	1.01	1.05

Table 24

Distribution of Mean-Square Fit Statistics: Grades 6–8 Listening

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 1 % = 2.6%
1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 2 % = 5.3%
0.5–1.5 “productive for measurement”	N = 37 % = 97.4%	N = 35 % = 92.1%
< 0.5 “less productive but not degrading”	N = 1 % = 2.6%	N = 0 % = 0%
Total	N = 38 % = 100%	N = 38 % = 100%

Table 25 presents the results of the Rasch analyses of the 38 Reading items for grade-level cluster 6-8, and Table 26 provides the summary of the fit statistics. The first four items are from Step 1; the latter items are for the Low, Mid, and High placement levels of Step 2 and for the Screener. The items with the label “COM” are common to both the Mid and High level forms. All infit and outfit statistics are considered to be productive for measurement of students’ English reading proficiency.

Table 25

Rasch Item Analysis: Grades 6–8 Reading

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_R68BC_LA_RestaurantReview_p2_Step_1	398	541	0.74	0.28	0.95	0.90
2.MODEL_R68BC_LA_RestaurantReview_p3_Step_1	213	541	0.39	1.91	1.04	1.05
3.MODEL_R68BC_LA_RestaurantReview_p4_Step_1	219	540	0.41	1.86	1.02	1.00
4.MODEL_R68BC_LA_RestaurantReview_p5_Step_1	111	541	0.21	2.94	1.14	1.33
5.MODEL_R68A_LA_CookingEggs_p1_Low	21	53	0.40	1.12	1.07	1.04
6.MODEL_R68A_LA_CookingEggs_p2_Low	28	53	0.53	0.51	0.89	0.88
7.MODEL_R68A_LA_CookingEggs_p3_Low	21	53	0.40	1.12	0.95	0.91
8.MODEL_R68A_MA_CoveringABox_p1_Low	46	53	0.87	-1.48	1.00	0.83
9.MODEL_R68A_MA_CoveringABox_p2_Low	25	53	0.47	0.77	0.95	0.91
10.MODEL_R68A_MA_CoveringABox_p3_Low	25	53	0.47	0.77	1.13	1.12
11.MODEL_R68A_SC_ConvectionCurrent_p1_Low	39	53	0.74	-0.52	0.85	0.75
12.MODEL_R68A_SC_ConvectionCurrent_p2_Low	35	53	0.66	-0.11	0.96	0.88
13.MODEL_R68A_SC_ConvectionCurrent_p3_Low	18	53	0.34	1.39	0.98	1.12
14.MODEL_R68B_MA_SchoolStore_p2_Mid	306	390	0.78	0.08	0.93	0.88
15.MODEL_R68B_MA_SchoolStore_p3_Mid	221	390	0.57	1.20	0.89	0.84
16.MODEL_R68B_MA_SchoolStore_p4_Mid	237	390	0.61	1.00	0.95	0.91
17.MODEL_R68B_SC_HowPlantsMakeFood_p2_Mid	340	390	0.87	-0.59	0.96	0.87
18.MODEL_R68B_SC_HowPlantsMakeFood_p3_Mid	180	390	0.46	1.67	1.00	1.03
19.MODEL_R68B_SC_HowPlantsMakeFood_p4_Mid	139	390	0.36	2.16	1.02	1.07
20.MODEL_R68B_SS_IndustrialRevolution_p2_Mid	341	390	0.87	-0.61	0.99	0.90
21.MODEL_R68B_SS_IndustrialRevolution_p3_Mid	75	390	0.19	3.09	0.89	0.86
22.MODEL_R68B_SS_IndustrialRevolution_p4_Mid	124	390	0.32	2.35	1.05	1.09
23.MODEL_R68C_LA_WinterSun_p3_COM	137	410	0.33	2.29	1.03	1.04
24.MODEL_R68C_LA_WinterSun_p4_COM	98	410	0.24	2.81	0.98	0.97
25.MODEL_R68C_LA_WinterSun_p5_COM	144	410	0.35	2.20	1.01	1.03
26.MODEL_R68C_MA_BookSale_p3_High	1	20	0.05	4.95	0.74	0.62
27.MODEL_R68C_MA_BookSale_p4_High	11	20	0.55	1.73	0.81	0.73
28.MODEL_R68C_MA_BookSale_p5_High	2	20	0.10	4.19	0.74	0.59
29.MODEL_R68C_SC_Pancakes_p3_High	1	20	0.05	4.95	1.26	1.26
30.MODEL_R68C_SC_Pancakes_p4_High	4	20	0.20	3.37	0.83	0.79
31.MODEL_R68C_SC_Pancakes_p5_High	3	20	0.15	3.72	1.13	1.12
32.MODEL_R68C_SS_PhotographyFirsts_p3_High	4	20	0.20	3.37	0.69	0.53
33.MODEL_R68C_SS_PhotographyFirsts_p4_High	7	20	0.35	2.58	0.74	0.67
34.MODEL_R68C_SS_PhotographyFirsts_p5_High	8	20	0.40	2.36	1.13	1.20
35.MODEL_R68C_MS_ChesapeakeBay_p3_Screener	23	74	0.31	2.34	1.00	1.00
36.MODEL_R68C_MS_ChesapeakeBay_p3_Screener	12	74	0.16	3.30	1.11	1.11
37.MODEL_R68C_MS_ChesapeakeBay_p4_Screener	20	74	0.27	2.62	1.17	1.31
38.MODEL_R68C_MS_ChesapeakeBay_p5_Screener	20	74	0.27	2.55	0.96	0.99

Table 26

Distribution of Mean-Square Fit Statistics: Grades 6–8 Reading

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 "distorting or degrading measurement"	N = 0 % = 0%	N = 0 % = 0%
1.5–2.0 "unproductive but not degrading"	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 "productive for measurement"	N = 38 % = 100%	N = 38 % = 100%
< 0.5 "less productive but not degrading"	N = 0 % = 0%	N = 0 % = 0%
Total	N = 38 % = 100%	N = 38 % = 100%

4.1.3.2. Rasch Analyses for the 9-12 Grade-level Cluster

Results of the Rasch item analysis for Grades 9–12 Speaking are reported in Table 27. The p-value and measure columns show that within the two parts of the assessment: a) each consecutive Speaking task was more difficult than the one before it, as entailed by the test design; b) more students answered tasks correctly in the first folder versus the second folder; and c) consistent with the *a priori* proficiency levels used during task development, fewer and fewer students responded correctly to test tasks as the proficiency level increased. Such patterns are consistent with the adaptive test design.

Table 28 summarizes the infit and outfit findings. For the infit, all ten items fall within the range considered by Linacre (2002) to be productive for measurement. For the outfit mean square statistics, five tasks have values between 0.5 and 1.5, one task is less than 0.5, and four tasks are greater than 2.0. For the four tasks with high outfit, the infit statistic falls within the productive range. Rasch users "routinely pay more attention to infit scores than outfit scores" (Bond & Fox, 2001, p. 43), as the infit statistics have been adjusted for outliers and are therefore less susceptible to inflation from inconsistent or unexpected responses, particularly for very easy or very difficult tasks (Linacre, 2002).

Table 27

Rasch Item Analysis: Grades 9–12 Speaking

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_S91_SI_FineArtsActivities_A_T1	499	526	0.95	-5.18	1.00	9.90
2.MODEL_S91_SI_FineArtsActivities_A_T2	425	526	0.81	-3.33	0.74	0.63
3.MODEL_S91_SI_FineArtsActivities_A_T3	334	526	0.64	-2.09	0.80	0.60
4.MODEL_S91_SI_FineArtsActivities_A_T4	182	526	0.35	-0.35	0.94	0.44
5.MODEL_S91_SI_FineArtsActivities_A_T5	106	526	0.20	0.74	0.97	0.53
6.MODEL_S91_LS_CarsAndAirPollution_B_T1	503	526	0.96	-5.38	1.12	1.31
7.MODEL_S91_LS_CarsAndAirPollution_B_T2	425	526	0.81	-3.33	0.87	2.82
8.MODEL_S91_LS_CarsAndAirPollution_B_T3	278	526	0.53	-1.45	0.94	1.02
9.MODEL_S91_LS_CarsAndAirPollution_B_T4	152	526	0.29	0.04	0.98	2.81
10.MODEL_S91_LS_CarsAndAirPollution_B_T5	70	526	0.13	1.48	1.16	6.59

Table 28

Distribution of Mean-Square Fit Statistics: Grades 9–12 Speaking

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 4 % = 40%
> 1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 10 % = 100%	N = 5 % = 50%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 1 % = 10%
Total	N = 10 % = 100%	N = 10 % = 100%

Table 29 presents the Rasch results for the 38 Listening items for the 9–12 grade-level cluster, and Table 30 summarizes the results. The first four items are from Step 1; the latter items are for the Low, Mid, and High placement levels of Step 2 and for the Screener. The items with the label “COM” are common to both the Mid and High level forms. No items were misfitting in terms of infit or outfit, indicating that all items are productive for measurement.

Table 29

Rasch Item Analysis: Grades 9–12 Listening

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_L91BC_LS_GroupBehavior_p2_Step_1	294	448	0.66	0.85	0.95	0.97
2.MODEL_L91BC_LS_GroupBehavior_p3_Step_1	249	448	0.56	1.35	1.02	1.17
3.MODEL_L91BC_LS_GroupBehavior_p4_Step_1	139	448	0.31	2.55	1.09	1.29
4.MODEL_L91BC_LS_GroupBehavior_p5_Step_1	236	448	0.53	1.49	1.24	1.29
5.MODEL_L91A_SI_SourcesofInformation_p1_Low	69	113	0.61	0.23	1.02	1.02
6.MODEL_L91A_SI_SourcesofInformation_p2_Low	53	113	0.47	0.90	0.88	0.85
7.MODEL_L91A_SI_SourcesofInformation_p3_Low	27	113	0.24	2.09	0.99	1.01
8.MODEL_L91A_LA_OneDayAfterSchool_p1_Low	51	113	0.45	0.98	0.88	0.87
9.MODEL_L91A_LA_OneDayAfterSchool_p2_Low	65	113	0.58	0.40	0.80	0.74
10.MODEL_L91A_LA_OneDayAfterSchool_p3_Low	57	113	0.50	0.73	0.88	0.85
11.MODEL_L91A_MA_CamillasPlant_p1_Low	46	113	0.41	1.19	1.00	0.98
12.MODEL_L91A_MA_CamillasPlant_p2_Low	35	113	0.31	1.68	0.97	0.89
13.MODEL_L91A_MA_CamillasPlant_p3_Low	55	113	0.49	0.81	1.05	1.06
14.MODEL_L91B_MA_LeesStore_p2_Mid	149	203	0.73	0.75	1.00	1.01
15.MODEL_L91B_MA_LeesStore_p3_Mid	64	203	0.32	2.69	0.99	1.02
16.MODEL_L91B_MA_LeesStore_p4_Mid	72	203	0.35	2.50	1.05	1.03
17.MODEL_L91B_SC_BlueCrabs_p2_Mid	143	203	0.70	0.91	0.94	0.92
18.MODEL_L91B_SC_BlueCrabs_p3_Mid	84	203	0.41	2.23	0.93	0.92
19.MODEL_L91B_SC_BlueCrabs_p4_Mid	79	203	0.39	2.34	1.01	1.01
20.MODEL_L91C_SS_PoliticalAlliances_p3_Mid	99	203	0.49	1.90	0.93	0.93
21.MODEL_L91C_SS_PoliticalAlliances_p4_Mid	87	203	0.43	2.16	0.98	0.99
22.MODEL_L91C_SS_PoliticalAlliances_p5_Mid	58	203	0.29	2.84	1.02	1.02
23.MODEL_L91C_LA_SeaStory_p3_COM	109	249	0.44	2.23	1.17	1.21
24.MODEL_L91C_LA_SeaStory_p4_COM	72	249	0.29	2.93	1.04	1.17
25.MODEL_L91C_LA_SeaStory_p5_COM	81	249	0.33	2.75	0.96	0.93
26.MODEL_L91C_MA_Statistics_p3_High	30	46	0.65	1.22	0.97	0.96
27.MODEL_L91C_MA_Statistics_p4_High	13	46	0.28	1.34	0.97	0.97
28.MODEL_L91C_MA_Statistics_p5_High	16	46	0.35	3.08	0.98	1.01
29.MODEL_L91C_SC_SingleCelledOrganisms_p3_High	35	46	0.76	2.98	0.84	0.77
30.MODEL_L91C_SC_SingleCelledOrganisms_p4_High	34	46	0.74	6.32	1.00	0.96
31.MODEL_L91C_SC_SingleCelledOrganisms_p5_High	15	46	0.33	3.75	0.98	0.96
32.MODEL_L91C_SS_EuropeanExplorers_p3_High	18	46	0.39	4.39	0.99	0.98
33.MODEL_L91C_SS_EuropeanExplorers_p4_High	15	46	0.33	3.39	0.97	1.04
34.MODEL_L91C_SS_EuropeanExplorers_p5_High	11	46	0.24	2.05	0.96	0.94
35.MODEL_L91C_MS_Balance_p3_Screener	55	83	0.66	0.84	1.14	1.08
36.MODEL_L91C_MS_Balance_p4_Screener	40	83	0.48	1.85	0.95	0.94
37.MODEL_L91C_MS_Balance_p5_Screener	16	83	0.19	3.26	0.91	0.75
38.MODEL_L91C_MS_Balance_p5_Screener	40	83	0.48	1.80	1.06	1.02

Table 30

Distribution of Mean-Square Fit Statistics: Grades 9–12 Listening

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 0 % = 0%
1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 38 % = 100%	N = 38 % = 100%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
Total	N = 38 % = 100%	N = 38 % = 100%

Table 31 presents the results of the Rasch analyses of the 38 Reading items for grade-level cluster 9-12, and Table 32 provides the summary of the fit statistics. The first four items are from Step 1; the latter items are for the Low, Mid, and High placement levels of Step 2 and for the Screener. The items with the label “COM” are common to both the Mid and High level forms. According to the infit statistic, all items fit the Rasch model well and are productive for measurement. For the outfit statistic, 37 items are productive for measurement, and one item provides distorting or degrading measurement information. This one item was answered correctly by a student who got nearly all other items incorrect.

Table 31

Rasch Item Analysis: Grades 9–12 Reading

Item Name	Score	Count	P-Value	Measure	IN.MSQ	OUT.MS
1.MODEL_R91BC_LA_NorthernSea_p2_Step_1	319	431	0.74	0.49	1.00	0.94
2.MODEL_R91BC_LA_NorthernSea_p3_Step_1	243	431	0.56	1.42	1.06	1.14
3.MODEL_R91BC_LA_NorthernSea_p4_Step_1	134	431	0.31	2.65	0.98	1.05
4.MODEL_R91BC_LA_NorthernSea_p5_Step_1	172	431	0.40	2.20	1.13	1.28
5.MODEL_R91A_LA_JuliaChild_p1_Low	62	103	0.60	0.41	0.87	0.81
6.MODEL_R91A_LA_JuliaChild_p2_Low	72	103	0.70	-0.07	0.93	0.91
7.MODEL_R91A_LA_JuliaChild_p3_Low	38	103	0.37	1.46	1.05	1.04
8.MODEL_R91A_MA_Polygons_p1_Low	84	103	0.82	-0.77	0.87	0.81
9.MODEL_R91A_MA_Polygons_p2_Low	57	103	0.55	0.63	0.95	0.93
10.MODEL_R91A_MA_Polygons_p3_Low	20	103	0.19	2.43	1.03	1.04
11.MODEL_R91A_SC_DietaryGuidelines_p1_Low	55	103	0.53	0.71	0.92	0.89
12.MODEL_R91A_SC_DietaryGuidelines_p2_Low	77	103	0.75	-0.33	0.92	0.87
13.MODEL_R91A_SC_DietaryGuidelines_p3_Low	12	103	0.12	3.07	0.98	2.24
14.MODEL_R91B_MA_AnglesinShapes_p2_Mid	83	225	0.37	2.52	0.93	0.93
15.MODEL_R91B_MA_AnglesinShapes_p3_Mid	112	225	0.50	1.94	0.87	0.86
16.MODEL_R91B_MA_AnglesinShapes_p4_Mid	75	225	0.33	2.70	0.95	0.92
17.MODEL_R91B_SC_ModelRockets_p2_Mid	77	225	0.34	2.65	1.18	1.28
18.MODEL_R91B_SC_ModelRockets_p3_Mid	57	225	0.25	3.12	1.06	1.13
19.MODEL_R91B_SC_ModelRockets_p4_Mid	105	225	0.47	2.08	1.10	1.12
20.MODEL_R91B_SS_AncientWriting_p2_Mid	111	225	0.49	1.96	0.98	0.97
21.MODEL_R91B_SS_AncientWriting_p3_Mid	115	225	0.51	1.88	0.98	0.96
22.MODEL_R91B_SS_AncientWriting_p4_Mid	108	225	0.48	2.02	0.94	0.93
23.MODEL_R91C_LA_ConservationistJoyAdamson_p3_COM	162	257	0.63	1.41	0.97	0.97
24.MODEL_R91C_LA_ConservationistJoyAdamson_p4_COM	138	257	0.54	1.84	1.02	1.02
25.MODEL_R91C_LA_ConservationistJoyAdamson_p5_COM	105	257	0.41	2.42	1.05	1.06
26.MODEL_R91C_MA_Perspective_p3_High	17	32	0.53	2.43	0.84	0.82
27.MODEL_R91C_MA_Perspective_p4_High	13	32	0.41	2.98	1.04	1.05
28.MODEL_R91C_MA_Perspective_p5_High	10	32	0.31	3.42	1.17	1.19
29.MODEL_R91C_SC_BacterialGrowth_p3_High	9	32	0.28	3.58	1.06	1.11
30.MODEL_R91C_SC_BacterialGrowth_p4_High	14	32	0.44	2.84	0.88	0.85
31.MODEL_R91C_SC_BacterialGrowth_p5_High	9	32	0.28	3.58	1.11	1.18
32.MODEL_R91C_SS_AncientCivilizations_p3_High	22	32	0.69	1.72	0.96	1.02
33.MODEL_R91C_SS_AncientCivilizations_p4_High	12	32	0.38	3.12	1.23	1.30
34.MODEL_R91C_SS_AncientCivilizations_p5_High	6	32	0.19	4.14	0.98	1.20
35.MODEL_R91C_MS_ElectricalCircuit_p3_Screener	24	68	0.35	2.65	1.02	1.10
36.MODEL_R91C_MS_ElectricalCircuit_p3_Screener	42	68	0.62	1.32	0.98	0.97
37.MODEL_R91C_MS_ElectricalCircuit_p4_Screener	17	68	0.25	3.30	0.96	1.28
38.MODEL_R91C_MS_ElectricalCircuit_p5_Screener	21	68	0.31	2.88	1.29	1.44

Table 32

Distribution of Mean-Square Fit Statistics: Grades 9–12 Reading

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 1 % = 2.6%
1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 38 % = 100%	N = 37 % = 97.4%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
Total	N = 38 % = 100%	N = 38 % = 100%

4.2. Results for MODEL Writing

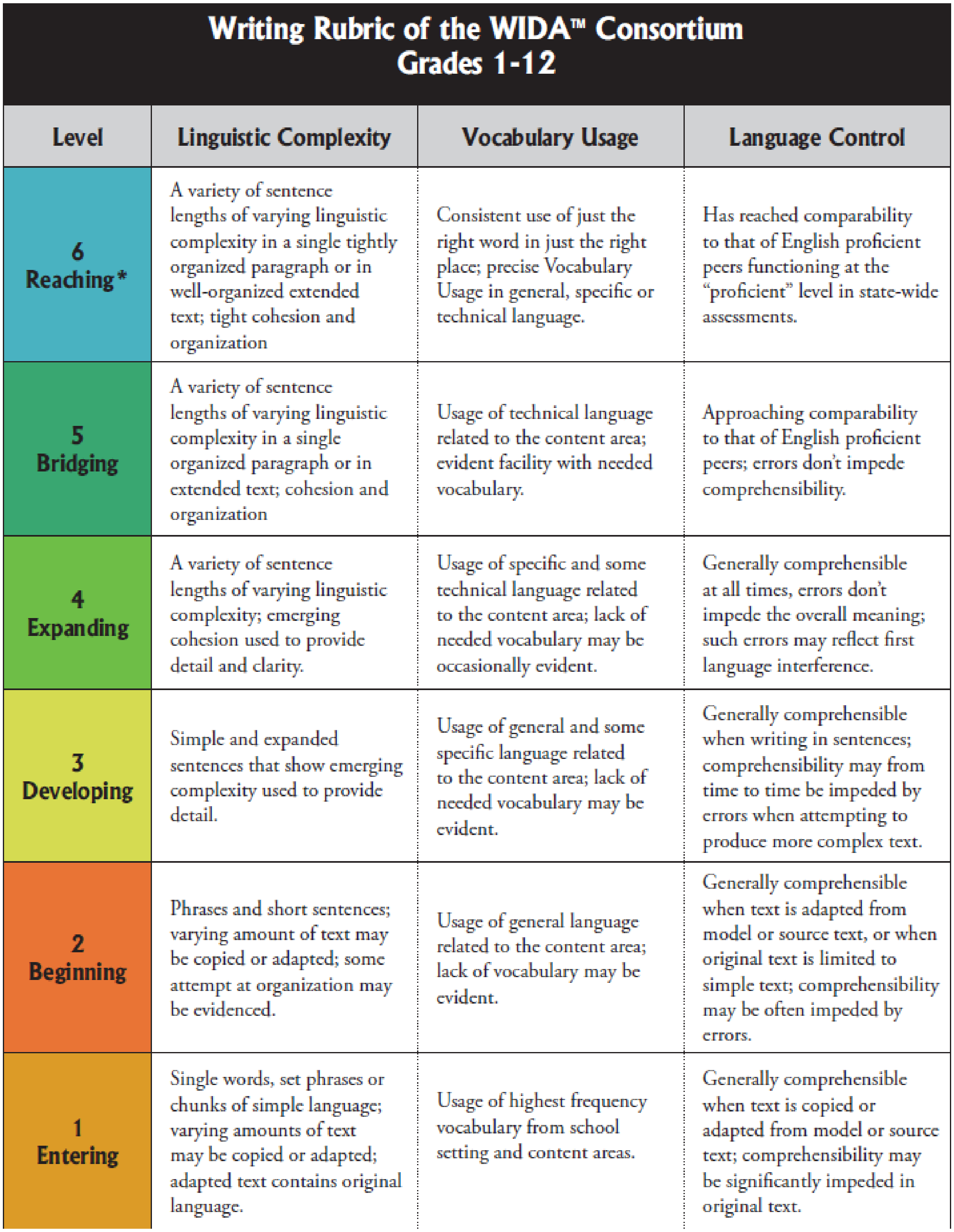
Students’ writing responses from the field test were analyzed to determine whether they could be accurately scored and to produce descriptive statistics for the tasks.

4.2.1. Scoring the Writing Responses

Scoring of the Writing responses was conducted in two phases. In Phase I, an internal CAL writing meeting was held to prepare for the operational scoring. In Phase II, an external writing meeting was conducted to score the students’ Writing responses operationally.

The internal CAL writing scoring meeting (Phase I) was held at CAL on January 5, 6, and 10, 2011. The facilitator was Dorry Kenyon, Director of CAL’s Language Testing Division, and panelists were CAL employees Stephanie Gibson, Daniel Ginsberg, Deepak Ebenezer, Tatyana Vdovina, and Abbe Spokane. The main goals of the meeting were to select sets of students’ writing samples to use in calibrating external raters to the WIDA Consortium’s Writing Rubric and to provide feedback to the MODEL Administrator Training team on materials that would eventually become part of the MODEL™ Training Tool Kit.

The WIDA Consortium’s Writing Rubric, shown in Figure 8, is a scoring guide in which a uniform set of criteria are used to interpret students’ Writing samples. The rubric was originally created to score the productive tasks in ACCESS and for its screener, the W-APT. The rubric catalogs the Performance Definitions for the six levels of English language proficiency and describes the three criteria, Linguistic Complexity, Vocabulary Usage, and Language Control, for each proficiency level.



Writing Rubric of the WIDA™ Consortium
Grades 1-12

Level	Linguistic Complexity	Vocabulary Usage	Language Control
6 Reaching*	A variety of sentence lengths of varying linguistic complexity in a single tightly organized paragraph or in well-organized extended text; tight cohesion and organization	Consistent use of just the right word in just the right place; precise Vocabulary Usage in general, specific or technical language.	Has reached comparability to that of English proficient peers functioning at the "proficient" level in state-wide assessments.
5 Bridging	A variety of sentence lengths of varying linguistic complexity in a single organized paragraph or in extended text; cohesion and organization	Usage of technical language related to the content area; evident facility with needed vocabulary.	Approaching comparability to that of English proficient peers; errors don't impede comprehensibility.
4 Expanding	A variety of sentence lengths of varying linguistic complexity; emerging cohesion used to provide detail and clarity.	Usage of specific and some technical language related to the content area; lack of needed vocabulary may be occasionally evident.	Generally comprehensible at all times, errors don't impede the overall meaning; such errors may reflect first language interference.
3 Developing	Simple and expanded sentences that show emerging complexity used to provide detail.	Usage of general and some specific language related to the content area; lack of needed vocabulary may be evident.	Generally comprehensible when writing in sentences; comprehensibility may from time to time be impeded by errors when attempting to produce more complex text.
2 Beginning	Phrases and short sentences; varying amount of text may be copied or adapted; some attempt at organization may be evidenced.	Usage of general language related to the content area; lack of vocabulary may be evident.	Generally comprehensible when text is adapted from model or source text, or when original text is limited to simple text; comprehensibility may be often impeded by errors.
1 Entering	Single words, set phrases or chunks of simple language; varying amounts of text may be copied or adapted; adapted text contains original language.	Usage of highest frequency vocabulary from school setting and content areas.	Generally comprehensible when text is copied or adapted from model or source text; comprehensibility may be significantly impeded in original text.

Figure 8: Writing Rubric of the WIDA Consortium

Source: *Understanding the WIDA English Language Proficiency Standards: A Resource Guide* (Gottlieb, Cranley, & Cammilleri, 2007)

The group of raters started with writing samples from Grades 6–8. In order to calibrate themselves to the WIDA Consortium's Writing Rubric and to ensure that samples could be accurately scored, the participants scored a set of 10 samples. Each person scored each writing sample and wrote his or her scores on colored Post-It notes. When participants met or exceeded expectations on Part A, Part B was administered. In these cases, Part A was scored after Part B because Part B is designed to elicit a response at a higher proficiency level. Raters used the rubric to decide which proficiency level, 1–6, best reflected the student's paper. This level was called the student's basic or solid score. If necessary, raters used a minus (-) or a plus (+) to indicate if the student's paper was weak in one feature (Linguistic Complexity, Vocabulary Usage, or Language Control) at that level (e.g., a weak 3, or 3-) or was exceptional in one feature at that level (e.g., a strong 3, or 3+).

The test administrator scored Part A if any one of the following was true: the student did not advance to Part B; the response in Part B was non-ratable (e.g., no response, incomprehensible, illegible, off-task, pictures, scribbles, refusal, not English); or the response in Part B received a score of 1 or 2 (anything less than a 3-). If both Part A and Part B were scored, both scores were recorded, but the higher of the two scores was considered to be the final writing score for the student.

Following a discussion of the samples and the rubric, the participants randomly selected and scored another 10 writing samples from the field test papers. After all participants understood how to properly score samples, approximately 50 samples for each of the two tasks were selected from the field test. Each person scored each sample independently. This scoring process was repeated for Grades 9–12. After the meeting, the raters' scores were manually typed by two CAL employees into their own separate spreadsheets in Microsoft Excel. Their data entry in the two spreadsheets was then compared, and any discrepancies were manually corrected.

To prepare for the external writing scoring meeting (Phase II), where all student writing assessments were subsequently scored, CAL employees organized the scores from the internal CAL writing scoring meeting from lowest to highest for each task. Papers that had the greatest rater agreement were selected to construct calibration sets for the external writing scorers. In total, CAL identified 20 calibration papers for each task (80 total) as the clearest examples of each score, across as many score points as possible.

For the external writing scoring meeting, CAL staff recruited external raters via advertisements in various distribution lists in the Washington, DC area. Qualified applicants had native or native-like proficiency of English, had a bachelor's degree or higher, had experience rating writing samples with a rubric, had previous work experience with middle and high school students and/or ELLs, and were available to work during certain days and hours. CAL recruited and accepted 13 external raters and 2 internal raters (i.e., CAL staff) to score papers for Grades 6–8 and 6 external raters and 1 internal rater to score papers for Grades 9–12. Selected participants were sent relevant training materials, including sections of the Test Administration Manual and instructional PowerPoint presentations. Participants were to study the materials on their own and to come to the meeting prepared to score writing samples.

The external writing scoring meetings were held over three days on February 2, 4, and 7, 2011 in Washington, DC. The primary goal of the meeting was to score the writing samples that were collected during the field test. Additionally, CAL staff were interested in observing the efficacy of the self-instructional materials and receiving user feedback on the Writing training materials. CAL employees Stephanie Gibson, Tatyana Vdovina, and Daniel Ginsberg facilitated the meeting.

At the beginning of the external Writing scoring meeting, raters scored approximately 10 prescored writing samples in order to gauge their present accuracy. After recording everyone's scores, participants engaged in a discussion about the ratings, with facilitators clarifying misunderstanding about the WIDA Consortium's Writing Rubric and the scoring procedures.

Next, to calibrate the raters and to expose them to the writing task that they would be scoring, each rater scored a calibration set of 10 pre-selected papers from the field test. Raters used the rubric to determine a basic or solid score ranging from 1–6 for Parts A and B. These scores could receive a plus (+) or minus (-) for strengths or weaknesses in Linguistic Complexity, Vocabulary Usage, or Language Control. Raters must achieve perfect or adjacent agreement with the scores assigned by CAL on at least 8 of the 10 pre selected papers to be qualified for scoring. Perfect and adjacent agreement shows the rate at which scores from two different raters were no more than one point score apart (e.g., if Rater 1 gives a score of 4, Rater 2 must give a score of 4 to be in perfect agreement or a score of 3 or 5 to be considered in adjacent agreement). Raters who did not meet this goal for the first calibration set were assigned a second calibration set of 10 papers for that task and again had the percentage of perfect and adjacent agreement computed. Raters who met the 80 percent criterion on the calibration sets were permitted to score the remainder of the field test papers for that task. One rater who did not meet the 80 percent criterion was not permitted to continue.

Operationally, each student's writing paper was scored by at least two different raters. Raters were randomly assigned sets of student papers as the first or second reader.

After the external writing scoring meeting, the raters' scores, which had been captured on scannable forms filled out by the raters themselves, were scanned by CAL research assistants. Each scannable form was scanned with Gravic® Remark Office OMR® software twice in case of scanner malfunction. The data were then cleaned in Microsoft Excel by comparing the data from the two scannings and manually reconciling any discrepancies. For ease of numerical analysis, these original scores were converted to raw scores ranging from 0–18, as shown in Table 33.

Table 33

Original Writing Scores and Their Corresponding Converted Raw Scores

Original Score	Converted Raw Score
NR	0
1-	1
1	2
1+	3
2-	4
2	5
2+	6
3-	7
3	8
3+	9
4-	10
4	11
4+	12
5-	13
5	14
5+	15
6-	16
6	17
6+	18

For each rater, researchers determined if the converted score for Part B or A was higher and kept that score as the final score for the student. The higher score was selected because Part A is a shorter, simpler task than Part B, so Part B would allow mid- and high-proficiency students to show more of their abilities and attain a higher score. For low-proficiency students, either Part A or Part B (if administered) might show their abilities best.

4.2.2. Descriptive Statistics for MODEL Writing

As described in Chapter 4.2.1, each student's writing paper was scored by at least two different raters. To account for the multiple scores assigned to each student as well as rater differences, the many-facets Rasch model (Facets software Version No. 3.58.0, Linacre, 2010) was used to derive the estimated raw score that a particular student's writing paper would have been obtained from a rater with average severity. Fair averages take into account rater variation in terms of harshness or leniency, so they are a better representation of student performance on the Writing tasks than simple averages.

A two-facet Rasch model was specified, which included a *student* facet and a *rater* facet:

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

where

P_{nik} = probability of person *n* on task *i* receiving a rating at level *k* on the rating scale
P_{nik-1} = probability of person *n* on task *i* receiving a rating at level *k-1* on the rating scale
B_n = ability of person *n*
D_i = difficulty of task *i*
α_j = severity of rater *j*
F_k = calibration of step *k* on the rating scale.

In this model, each Writing task is characterized by a *difficulty*, D_i, each Student by *ability*, B_n, and each rater by a level of *severity*, α_j. The log odds formulation places the parameters on a common scale of log odds units or logit. Facets used the scores that raters awarded to Students' papers to estimate the individual Student abilities and rater severity levels.

The frequency distribution, mean, and standard deviation of the rounded Fair Averages for Writing Task 1 for Grades 6–8 are shown in Table 34, and the statistics for Task 2 are shown in Table 35.

Table 34

Descriptive Statistics for Writing Fair Averages: Grades 6–8 Writing Task 1

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	0	0.0	0.0
1	0	0.0	0.0
2	5	1.6	1.6
3	3	0.9	2.5
4	7	2.2	4.7
5	56	17.7	22.4
6	80	25.2	47.6
7	79	24.9	72.6
8	59	18.6	91.2
9	21	6.6	97.8
10	5	1.6	99.4
11	2	0.6	100.0
12	0	0.0	100.0
13	0	0.0	100.0
14	0	0.0	100.0
15	0	0.0	100.0
16	0	0.0	100.0
17	0	0.0	100.0
18	0	0.0	100.0
Total	317	100.0	
Mean	6.63		
Standard Deviation	1.48		

Table 35

Descriptive Statistics for Writing Fair Averages: Grades 6–8 Writing Task 2

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	0	0.0	0.0
1	1	0.2	0.2
2	8	2.0	2.2
3	2	0.5	2.7
4	7	1.7	4.5
5	37	9.2	13.6
6	83	20.6	34.2
7	70	17.4	51.6
8	102	25.3	76.9
9	47	11.7	88.6
10	26	6.5	95.0
11	13	3.2	98.3
12	4	1.0	99.3
13	2	0.5	99.8
14	0	0.0	99.8
15	0	0.0	99.8
16	1	0.2	100.0
17	0	0.0	100.0
18	0	0.0	100.0
Total	403	100.0	
Mean	7.33		
Standard Deviation	1.89		

Table 36 and Table 37 show the frequency distributions, means, and standard deviations of the rounded Fair Averages for each Writing task for Grades 9–12.

Table 36

Descriptive Statistics for Writing Fair Averages: Grades 9–12 Writing Task 1

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	3	1.0	1.0
1	5	1.6	2.6
2	22	7.1	9.7
3	25	8.1	17.9
4	43	14.0	31.8
5	72	23.4	55.2
6	70	22.7	77.9
7	41	13.3	91.2
8	7	2.3	93.5
9	12	3.9	97.4
10	3	1.0	98.4
11	3	1.0	99.4
12	1	0.3	99.7
13	0	0.0	99.7
14	0	0.0	99.7
15	1	0.3	100.0
16	0	0.0	100.0
17	0	0.0	100.0
18	0	0.0	100.0
Total	308	100.0	
Mean	5.24		
Standard Deviation	2.03		

Table 37

Descriptive Statistics for Writing Fair Average: Grades 9–12 Writing Task 2

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	1	0.5	0.5
1	2	1.0	1.4
2	24	11.6	13.0
3	15	7.2	20.3
4	20	9.7	30.0
5	21	10.1	40.1
6	52	25.1	65.2
7	34	16.4	81.6
8	22	10.6	92.3
9	9	4.3	96.6
10	4	1.9	98.6
11	2	1.0	99.5
12	1	0.5	100.0
13	0	0.0	100.0
14	0	0.0	100.0
15	0	0.0	100.0
16	0	0.0	100.0
17	0	0.0	100.0
18	0	0.0	100.0
Total	207	100.0	
Mean	5.54		
Standard Deviation	2.25		

4.3. Results for MODEL Screener

4.3.1. Rasch Analyses for MODEL Screener Listening and Reading Sections

Table 38 presents the results of the Rasch analyses of the eight Screener Listening items for Grades 6–8, and Table 39 summarizes the findings. All items have good infit and outfit mean square statistics according to the guidelines provided by Linacre (2002). These items fit the Rasch model well and are productive for measurement. The fit statistics differ from those presented in Chapter 4.1 because the statistics presented here are for Screener students only.

Table 38

Rasch Item Analysis: Grades 6–8 Listening Screener

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_L68BC_LA_TheTraveler_p2_Step_1	61	77	0.79	0.42	1.17	1.01
2.MODEL_L68BC_LA_TheTraveler_p3_Step_1	36	77	0.47	1.91	0.95	0.91
3.MODEL_L68BC_LA_TheTraveler_p4_Step_1	46	77	0.60	1.82	0.98	0.92
4.MODEL_L68BC_LA_TheTraveler_p5_Step_1	52	77	0.68	1.00	1.00	1.04
5.MODEL_L68C_MS_HangingScale_p3_Screener	50	77	0.65	1.19	0.92	0.85
6.MODEL_L68C_MS_HangingScale_p4_Screener	7	77	0.09	4.19	1.12	1.41
7.MODEL_L68C_MS_HangingScale_p5_Screener	21	77	0.27	3.00	0.85	0.75
8.MODEL_L68C_MS_HangingScale_p5_Screener	21	77	0.27	3.07	1.13	1.05

Table 39

Distribution of Mean-Square Fit Statistics: Grades 6–8 Listening Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 0 % = 0%
> 1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 8 % = 100%	N = 8 % = 100%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
Total	N = 8 % = 100%	N = 8 % = 100%

Table 40 presents the results of the Rasch analyses on the eight Screener Reading items for Grades 6–8, and Table 41 summarizes the findings. All items have good infit and outfit mean square statistics according to the guidelines provided by Linacre (2002). These items fit the Rasch model well and are productive for measurement. The fit statistics differ from those presented in Chapter 4.1 because the statistics presented here are for Screener students only.

Table 40

Rasch Item Analysis: Grades 6–8 Reading Screener

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_R68BC_LA_RestaurantReview_p2_Step_1	48	74	0.65	0.28	0.99	0.98
2.MODEL_R68BC_LA_RestaurantReview_p3_Step_1	24	74	0.32	1.91	0.99	0.92
3.MODEL_R68BC_LA_RestaurantReview_p4_Step_1	29	74	0.39	1.86	0.80	0.75
4.MODEL_R68BC_LA_RestaurantReview_p5_Step_1	19	74	0.26	2.94	1.00	1.05
5.MODEL_R68C_MS_ChesapeakeBay_p3_Screener	23	74	0.31	2.34	1.00	1.00
6.MODEL_R68C_MS_ChesapeakeBay_p3_Screener	12	74	0.16	3.30	1.11	1.11
7.MODEL_R68C_MS_ChesapeakeBay_p4_Screener	20	74	0.27	2.62	1.17	1.31
8.MODEL_R68C_MS_ChesapeakeBay_p5_Screener	20	74	0.27	2.55	0.96	0.99

Table 41

Distribution of Mean-Square Fit Statistics: Grades 6–8 Reading Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 0 % = 0%
> 1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 8 % = 100%	N = 8 % = 100%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
Total	N = 8 % = 100%	N = 8 % = 100%

The Rasch results for the Screener Speaking section and the Facets results for the Screener Writing section for Grades 6–8 are the same as those presented in Chapter 4 (see Table 21 and Table 22 for Speaking results and Table 34 and Table 35 for Writing results), because all Speaking and Writing tasks from the MODEL were also included in MODEL Screener.

Table 42 presents the results of the Rasch analyses on the eight Screener Listening items for Grades 9–12 and summarizes the findings. All items have good infit and outfit mean square statistics according to the guidelines provided by Linacre (2002). These items fit the Rasch model well and are productive for measurement. The fit statistics differ from those presented in Chapter 4.1 because the statistics presented here are for Screener students only.

Table 42

Rasch Item Analysis: Grades 9–12 Listening Screener

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_L91BC_LS_GroupBehavior_p2_Step_1	56	83	0.67	0.85	0.89	0.80
2.MODEL_L91BC_LS_GroupBehavior_p3_Step_1	51	83	0.61	1.35	0.88	0.83
3.MODEL_L91BC_LS_GroupBehavior_p4_Step_1	27	83	0.33	2.55	1.02	0.93
4.MODEL_L91BC_LS_GroupBehavior_p5_Step_1	53	83	0.64	1.49	1.20	1.29
5.MODEL_L91C_MS_Balance_p3_Screener	55	83	0.66	0.84	1.14	1.08
6.MODEL_L91C_MS_Balance_p4_Screener	40	83	0.48	1.85	0.95	0.94
7.MODEL_L91C_MS_Balance_p5_Screener	16	83	0.19	3.26	0.91	0.75
8.MODEL_L91C_MS_Balance_p5_Screener	40	83	0.48	1.8	1.06	1.02

Table 43

Distribution of Mean-Square Fit Statistics: Grades 9–12 Listening Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 0 % = 0%
> 1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 8 % = 100%	N = 8 % = 100%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
Total	N = 8 % = 100%	N = 8 % = 100%

Table 44 presents the results of the Rasch analyses on the eight Screener Reading items for Grades 9–12, and Table 45 summarizes the findings. All items have good infit and outfit mean square statistics according to the guidelines provided by Linacre (2002). These items fit the Rasch model well and are productive for measurement. The fit statistics differ from those presented in Chapter 4.1 because the statistics presented here are for Screener students only.

Table 44

Rasch Item Analysis: Grades 9–12 Reading Screener

Item Name	Score	Count	P-value	Measure	IN.MSQ	OUT.MSQ
1.MODEL_R91BC_LA_NorthernSea_p2_Step_1	49	68	0.72	0.49	1.04	0.99
2.MODEL_R91BC_LA_NorthernSea_p3_Step_1	42	68	0.62	1.42	0.84	1.02
3.MODEL_R91BC_LA_NorthernSea_p4_Step_1	19	68	0.28	2.65	0.69	0.52
4.MODEL_R91BC_LA_NorthernSea_p5_Step_1	29	68	0.43	2.20	1.09	1.08
5.MODEL_R91C_MS_ElectricalCircuit_p3_Screener	24	68	0.35	2.65	1.02	1.10
6.MODEL_R91C_MS_ElectricalCircuit_p3_Screener	42	68	0.62	1.32	0.98	0.97
7.MODEL_R91C_MS_ElectricalCircuit_p4_Screener	17	68	0.25	3.30	0.96	1.28
8.MODEL_R91C_MS_ElectricalCircuit_p5_Screener	21	68	0.31	2.88	1.29	1.44

Table 45

Distribution of Mean-Square Fit Statistics: Grades 9–12 Reading Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 “distorting or degrading measurement”	N = 0 % = 0%	N = 0 % = 0%
> 1.5–2.0 “unproductive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
0.5–1.5 “productive for measurement”	N = 8 % = 100%	N = 8 % = 100%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 0 % = 0%
Total	N = 8 % = 100%	N = 8 % = 100%

The Rasch results for the MODEL Screener Speaking section and the Facets results for the MODEL Screener Writing section for Grades 9–12 are the same as those presented in Chapter 4 (see Table 27 and Table 28 for MODEL Speaking results and Table 36 and Table 37 for MODEL Writing results), as all Speaking and Writing tasks for the MODEL were included on MODEL Screener.

4.3.2. MODEL Screener Descriptive Statistics

Table 46 presents the descriptive statistics for the Listening and Reading sections of MODEL Screener for Grades 6–8. Also included are the descriptive statistics for the Speaking and Writing sections, which are identical to those from the Speaking and Writing in MODEL because MODEL Screener included all Speaking and Writing tasks. Note that the Writing statistics were computed based on students’ Fair Averages from the Facets analysis (see Chapter 4.2.2 for more information).

Table 46

Descriptive Statistics for MODEL Screener by Domain and Task for Grades 6–8

Domain		No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking		10	728	0	10	7.06	2.22
Listening		8	106	1	8	3.82	1.46
Reading		8	102	1	8	2.76	1.33
Writing	Task 1	-	317	2	11	6.63	1.48
	Task 2	-	403	1	16	7.33	1.89

Table 47 presents the descriptive statistics for the Speaking, Listening, Reading, and Writing sections of the Screener for Grades 9–12. The descriptive statistics for the Speaking and Writing sections are identical to those from the Speaking and Writing in MODEL because MODEL Screener included all Speaking and Writing tasks.

Table 47

Descriptive Statistics for MODEL Screener by Domain and Task for Grades 9–12

Domain		No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking		10	526	0	10	5.65	2.65
Listening		8	93	1	8	4.04	1.57
Reading		8	77	1	7	3.55	1.63
Writing	Task 1	-	308	0	15	5.24	2.03
	Task 2	-	207	0	12	5.54	2.25

5. Linking WIDA MODEL to WIDA ELP Levels

This chapter presents the procedure for linking the WIDA MODEL series for Grades 6-8 and 9-12 to the WIDA ELP levels. As discussed in Chapter 1 of this report, in order to make the scores on WIDA MODEL more intelligible to educators, students' performances on WIDA MODEL are interpreted in terms of WIDA's ELP levels. Score interpretations are presented in the form of lookup tables that show the WIDA ELP level scores that correspond with students' raw scores and scale scores at each grade and for each domain. This chapter of the report explains the linking studies that were conducted to link WIDA MODEL scores to WIDA ELP levels and how the lookup tables were derived from these studies. The WIDA MODEL lookup tables can be found in the appendices of the *WIDA MODEL™ Test Administration Manual* for Grades 6–8 and Grades 9–12 (MetriTech & CAL, 2011).

WIDA MODEL was developed to measure the same WIDA ELP Standards as ACCESS, for which a Standard Setting Study was held in Madison, WI from April 20–27, 2005. The ACCESS Standard Setting Study used the WIDA ELP Standards together with empirical information from field test data to determine the relationship between student performances on the four domains and the language proficiency levels defined by the WIDA ELP Standards. More details about the ACCESS Standard Setting Study are in the report *Development and Field Test of ACCESS for ELLs®* (Kenyon, 2006).

For Listening and Reading, once WIDA MODEL scores were linked to the ACCESS scales, performances on WIDA MODEL could be interpreted in terms of the WIDA ELP levels. For Writing and Speaking, qualitative interpretations of performances on WIDA MODEL were used to directly relate performances on WIDA MODEL to the WIDA ELP levels. Details about the methodologies used to determine how performances on WIDA MODEL were linked and interpreted in terms of the WIDA ELP levels are presented in the following sections of this chapter.

5.1. Linking Listening and Reading Scores on WIDA MODEL and ACCESS

Because field test participants took both the WIDA MODEL field test and ACCESS Series 202 test, a common-person design was used to establish the link between the two assessments for the Listening and Reading domains. A concurrent calibration procedure was used to place the WIDA MODEL Listening and Reading scores on the same scales as the ACCESS Listening and Reading scores. In this procedure, the item difficulties of the WIDA MODEL items were estimated while the item difficulties of the ACCESS test items were fixed using Winsteps (Linacre, 2012b). Student test data from both WIDA MODEL and ACCESS were used as input for the estimation. Through this concurrent calibration procedure, the item difficulty parameters of the WIDA MODEL Listening and Reading items were placed on the same scale as the

ACCESS Series 202 Listening and Reading items, allowing the WIDA MODEL scores to be placed on the same scale as the ACCESS scores.

5.1.1. Listening and Reading Scale Score Adjustments

After the concurrent calibrations were completed for Listening and Reading, initial sets of raw score to scale score conversions were produced by grade and placement level and adjusted using the procedure described in this section. The adjustment was done by subtracting the standard error of measurement value from the initial raw score to scale score conversions. This is the same approach used for the W-APT and is done to compensate for the relatively large standard error of measurement on the shorter WIDA MODEL test forms. Adjusting the scale score down helps to ensure that students' abilities will not be overestimated, which could potentially lead to schools placing them in classes or programs for which they are not prepared. Proficiency level scores are interpretations of a student's scale score in terms of the WIDA ELP Standards, so the adjustment made to the WIDA MODEL scale score ensures that student proficiency levels for the domains of Listening and Reading are at least the proficiency level indicated by the assessment.

5.2. Linking Writing and Speaking Scores on WIDA MODEL and ACCESS

For Writing and Speaking, qualitative interpretations of performances on WIDA MODEL were used to directly relate performances on WIDA MODEL to the WIDA ELP levels.

5.2.1. Writing

To link scores on the WIDA MODEL Writing section to scores on the ACCESS Writing section, an expert panel of four CAL staff members examined 14 papers collected from the WIDA MODEL Grade 6-8 Writing field test and 11 papers collected from the Grade 9-12 Writing field test. The WIDA MODEL papers had already been rated, and each one corresponded to a different raw score point (e.g., 1-, 1, 1+, etc.). The expert panelists independently evaluated each WIDA MODEL paper by comparing it to a set of portfolios that had been used in standard setting for ACCESS (Kenyon, 2006) and had been assigned scale scores based on the WIDA ELP performance level descriptors. Panelists identified the ACCESS portfolio that most resembled each WIDA MODEL paper and, based on the score of the ACCESS portfolio, assigned scale scores to the WIDA MODEL writing samples. They then discussed their results together and were given a chance to adjust their judgments in a second round. The panelists' scores after the second round were averaged to determine the scale score corresponding to each raw score point.

Table 48 and Table 49 show the results of the study for Grades 6–8 and 9–12, respectively. The raw score is the value given to the WIDA MODEL Writing sample by the rater. The *a priori* proficiency levels transform the raw scores in the first column to values that are consistent with the proficiency levels of the WIDA MPIs. The scale scores are reported in the third column (procedures for deriving the scale score were described in the last paragraph), and the corresponding proficiency levels for each grade are shown in the last three columns. The grade-level proficiency scores are based on the interpretations of each assigned scale score in ACCESS. For example, a rater-assigned raw score of 2- is interpreted as an *a priori* proficiency level score of 2.2. During the expert panel, the panelists determined that the 2- writing sample they reviewed was the equivalent of scale score of 306 on ACCESS; using the ACCESS conversion, a scale score of 306 corresponds to a proficiency level score of 1.8 for Grade 8, 1.9 for Grade 7, and 2.3 for Grade 6. (Note that no portfolios had raw scores of "5," "6-," "6," or "6+" for Grades 6–8, and no portfolios had raw scores of "4+," "5-," "5," "6-," "6," or "6+" for Grades 9–12.).

Table 48

Writing Scale Scores and Proficiency Levels: Grades 6–8

Raw Score	*A priori* Proficiency Level	Scale Score	Grade 8 Proficiency Level	Grade 7 Proficiency Level	Grade 6 Proficiency Level
1-	1.2	293	1.7	1.8	1.9
1	1.5	297	1.7	1.9	1.9
1+	1.8	301	1.8	1.9	2.1
2-	2.2	306	1.8	1.9	2.3
2	2.5	314	1.9	2.2	2.5
2+	2.8	325	2.2	2.6	2.9
3-	3.2	340	2.8	3.0	3.4
3	3.5	369	3.6	3.9	4.2
3+	3.8	373	3.8	4.0	4.4
4-	4.2	379	3.9	4.3	4.6
4	4.5	382	4.0	4.4	4.7
4+	4.8	385	4.2	4.5	4.8
5-	5.2	396	4.6	4.9	5.3
5	5.5	n/a	n/a	n/a	n/a
5+	5.8	400	4.7	5.1	5.5
6-	6.0	n/a	n/a	n/a	n/a
6	6.0	n/a	n/a	n/a	n/a
6+	6.0	n/a	n/a	n/a	n/a

Table 49

Writing Scale Scores and Proficiency Levels: Grades 9–12

Raw Score	*A priori* Proficiency Level	Scale Score	Grade 12 Proficiency Level	Grade 11 Proficiency Level	Grade 10 Proficiency Level	Grade 9 Proficiency Level
1-	1.2	335	1.8	1.9	1.9	2.3
1	1.5	337	1.8	1.9	2.0	2.4
1+	1.8	343	1.9	1.9	2.3	2.6
2-	2.2	351	1.9	2.3	2.6	2.9
2	2.5	360	2.3	2.6	2.9	3.1
2+	2.8	371	2.8	3.0	3.2	3.4
3-	3.2	380	3.1	3.3	3.5	3.8
3	3.5	384	3.2	3.4	3.6	3.8
3+	3.8	387	3.3	3.5	3.7	3.9
4-	4.2	398	3.7	3.8	4.0	4.4
4	4.5	407	3.9	4.1	4.4	4.7
4+	4.8	n/a	n/a	n/a	n/a	n/a
5-	5.2	n/a	n/a	n/a	n/a	n/a
5	5.5	n/a	n/a	n/a	n/a	n/a
5+	5.8	444	5.6	5.9	6.0	6.0
6-	6.0	n/a	n/a	n/a	n/a	n/a
6	6.0	n/a	n/a	n/a	n/a	n/a
6+	6.0	n/a	n/a	n/a	n/a	n/a

After the Writing expert panel, the proficiency level for the highest grade level in each cluster was compared to the *a priori* value and appropriately adjusted to create the final lookup tables, which are presented in Table 50 and Table 51 below. Note that information for Grade 9 is included in the lookup table for Grade 6-8; this is because students in the first semester of Grade 9 are administered the Grade 6-8 test form, as described in Section 1.3.1.

Table 50

Writing Lookup Table: Grade 6–8 Form

	Grade 6		Grade 7		Grade 8		Grade 9	
Raw Score	Scale Score	Proficiency Level Score	Scale Score	Proficiency Level Score	Scale Score	Proficiency Level Score	Scale Score	Proficiency Level Score
0	233	1.0	239	1.0	245	1.0	251	1.0
1-	245	1.2	245	1.1	245	1.0	251	1.0
1	270	1.6	270	1.5	270	1.3	270	1.3
1+	291	1.9	291	1.8	291	1.6	291	1.5
2-	318	2.7	318	2.3	318	2.0	318	1.9
2	328	2.9	328	2.7	328	2.3	328	2.0
2+	336	3.2	336	2.9	336	2.6	336	2.3
3-	348	3.6	348	3.3	348	3.0	348	2.8
3	359	3.9	359	3.6	359	3.3	359	3.1
3+	368	4.2	368	3.9	368	3.6	368	3.4
4-	381	4.7	381	4.4	381	4.0	381	3.8
4	390	4.9	390	4.7	390	4.3	390	4.0
4+	397	5.3	397	4.9	397	4.6	397	4.3
5-	408	5.9	408	5.5	408	5.0	408	4.8
5	412	6.0	414	5.8	414	5.3	414	4.9
5+	412	6.0	420	6.0	420	5.6	420	5.3
6-	412	6.0	420	6.0	428	6.0	428	5.7
6	412	6.0	420	6.0	428	6.0	435	6.0
6+	412	6.0	420	6.0	428	6.0	435	6.0

Table 51

Writing Lookup Table: Grade 9–12 Form

	Grade 9		Grade 10		Grade 11		Grade 12	
Raw Score	Scale Score	Proficiency Level Score	Scale Score	Proficiency Level Score	Scale Score	Proficiency Level Score	Scale Score	Proficiency Level Score
0	251	1.0	257	1.0	263	1.0	269	1.0
1-	269	1.2	269	1.2	269	1.1	269	1.0
1	297	1.6	297	1.5	297	1.4	297	1.3
1+	322	1.9	322	1.8	322	1.7	322	1.6
2-	352	2.9	352	2.6	352	2.3	352	2.0
2	360	3.1	360	2.9	360	2.6	360	2.3
2+	367	3.3	367	3.1	367	2.9	367	2.6
3-	377	3.7	377	3.4	377	3.2	377	3.0
3	388	3.9	388	3.8	388	3.5	388	3.3
3+	397	4.3	397	4.0	397	3.8	397	3.6
4-	410	4.8	410	4.5	410	4.3	410	4.0
4	418	5.2	418	4.9	418	4.6	418	4.3
4+	424	5.5	424	5.1	424	4.9	424	4.6
5-	434	5.9	434	5.7	434	5.3	434	5.0
5	435	6.0	439	5.9	439	5.6	439	5.3
5+	435	6.0	441	6.0	445	5.9	445	5.6
6-	435	6.0	441	6.0	447	6.0	452	6.0
6	435	6.0	441	6.0	447	6.0	452	6.0
6+	435	6.0	441	6.0	447	6.0	452	6.0

5.2.2. Speaking

For the WIDA MODEL Speaking section, the scores were interpreted using the same procedure as the Speaking proficiency scores for ACCESS (Kenyon, 2006). Because the tasks for ACCESS Speaking were written to elicit speech samples at specific, progressively higher proficiency levels, the Standard Setting Panel for ACCESS decided that an examinee had to respond successfully to all prompts at a given proficiency level and below in order to be rated at that level. Because the ACCESS Speaking section has three tasks designed to elicit speech at Proficiency Level 1 and three tasks at Proficiency Level 2, an examinee is required to respond successfully to at least six tasks before being rated at Proficiency Level 2. In addition, the Standard Setting Panel determined that a perfect score should be rated at Proficiency Level 6 on ACCESS. In the case of WIDA MODEL, with only two folders designed to elicit speech at each of the five proficiency levels, a raw score of 4 was required for examinees to be rated at Proficiency Level 2. Table 52 and Table 53 show the WIDA MODEL Speaking scale score associated with each raw score, as well as the corresponding proficiency level by grade, starting with the highest grade in each cluster.

Table 52

Speaking Scale Scores and Proficiency Levels: Grades 6–8

Raw Score	Scale Score	Proficiency Level		
		Grade 8	Grade 7	Grade 6
0	180	1.0	1.0	1.0
1	221	1.3	1.3	1.3
2	246	1.5	1.5	1.5
3	274	1.7	1.7	1.7
4	317	2.0	2.1	2.3
5	330	2.5	2.6	2.8
6	344	3.0	3.2	3.5
7	352	3.5	3.7	3.9
8	361	4.0	4.1	4.3
9	384	5.0	5.2	5.4
10	404	6.0	6.0	6.0

Table 53

Speaking Scale Scores and Proficiency Levels: Grades 9–12

Raw Score	Scale Score	Proficiency Level			
		Grade 12	Grade 11	Grade 10	Grade 9
0	184	1.0	1.0	1.0	1.0
1	225	1.3	1.3	1.3	1.3
2	252	1.5	1.5	1.5	1.5
3	280	1.7	1.7	1.7	1.7
4	323	2.0	2.0	2.1	2.1
5	340	2.5	2.6	2.7	2.8
6	357	3.0	3.1	3.3	3.6
7	370	3.5	3.7	3.9	4.2
8	384	4.0	4.3	4.6	4.9
9	405	5.0	5.4	5.7	5.9
10	421	6.0	6.0	6.0	6.0

5.3. Truncating and Capping of Scale Scores

As is done for ACCESS, scale scores for all domains were adjusted so that a raw score of 0 was assigned the lowest proficiency level of 1.0. In addition, domain scores were capped at the Proficiency Level 5/6 cut score level since the WIDA ELP Standards do not contain Performance Definitions for Proficiency Level 6.

6. Validity

6.1. Validity Argument

“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” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [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 as a way to assess whether there is evidence that supports the appropriateness and adequacy of the interpretations and decisions made about test takers on the basis of their performance on a test. The Assessment Use Argument (Bachman & Palmer, 2010) is a conceptual validation framework consisting of a series of inferences that link the test taker’s performance to a claim along with evidence to support the claim. Following Bachman & Palmer (2010), this chapter presents a condensed Assessment Use Argument to link students’ scores on WIDA MODEL to an interpretation of their English language proficiency. The following sections of this report describe two claims that were investigated and the data and evidence that were collected to support the claims.

The validity evidence provided in this chapter applies to both MODEL and MODEL Screener.

6.2. Claim 1: Collection of Consistent Test Taker Data

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, including different test items, different test administrations, different times, or different raters. Analysis of test reliability provides information about the likelihood that students would receive the same score on the test over repeated test administrations.

Claim 1: Test takers’ performances on WIDA MODEL are consistent, including across different aspects of assessment conditions.

6.2.1. Evidence from Test Administration Procedures

WIDA has attempted to address environmental factors by specifying 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: the *WIDA MODEL™ Test Administration Manual* (MetriTech & CAL, 2011), which details how to prepare for, administer, score, and interpret scores on MODEL™; the *WIDA MODEL™ Test Administration Training DVD* (WIDA, 2011c), which covers general information on the structure of WIDA MODEL, includes commentary from the test developers, and shows scenes demonstrating test administration; and the *WIDA MODEL™ Assessment Training Toolkit CD-ROM* (WIDA, 2011b), which introduces various presentations and resources (i.e., PDFs, Excel workbooks, and PowerPoint presentations) available for people preparing to administer WIDA MODEL.

6.2.2. Evidence from Test Development Procedures

Additional evidence was provided by a series of qualitative evaluations of WIDA MODEL test content during the WIDA MODEL test development process. The bias and content review (Chapter 2.4), the bias and sensitivity review (Chapter 2.5), and the international perspectives panel (Chapter 2.6) were conducted with content experts to help ensure that items were appropriate and universal to people of different ethnic backgrounds and that items did not contain cultural bias or sensitive topics. The knowledge, expertise, and professional judgments of the experts ultimately ensured that WIDA MODEL provides all test takers with comparable opportunities to demonstrate their English language proficiency.

6.2.3. Evidence from Reliability of Step 2 Placement

The placement rules were designed such that the majority of the students are placed in the Mid level and only the students who did very poorly or very well in Step 1 Listening and Reading are placed in the Low and High level. Placing students at the proper test level allows for a more accurate measurement of students' abilities and reduces measurement error due to floor or ceiling effects. The descriptive statistics for the Step 2 items were analyzed to examine whether there are potential floor or ceiling effect. The presence of floor and ceiling help to evaluate the effectiveness of the placement rules. A **floor effect** occurs when large numbers of students answer all items incorrectly. This is a particular concern for the Mid and High placement level since these students may have been placed in a level that is overly challenging for them. For Low placement level, however, since some students placed at this level may have very limited experiences in English, a floor effect, if observed may not be a serious concern. A **ceiling effect** occurs when large numbers of students answer all items correctly; this is a particular concern for the Low and Mid placement level since these students might have demonstrated a higher proficiency if they had been placed at a higher placement level. Since the High placement level is the highest placement possible, a ceiling effect, if observed, is not a major concern in this context. Data used in the placement rule analyses are before the outliers were removed (see Table 16 and Table 17 for the number of students used in this analysis).

The placement algorithm for the Listening section, first mentioned in Chapter 1.3.3 of this report, is as follows: All students complete Listening Step 1; that score is combined with the Speaking score to determine a placement of Low, Mid, or High for Listening Step 2. For Grades 6–8, no students in any placement level answered all Step 2 items incorrectly, and only one student in the Mid placement level answered all Step 2 items correctly. The results do not suggest reason for concern about floor or ceiling effects for 6-8 Listening. For Grades 9–12, only one student in the Low placement level incorrectly answered all Step 2 items. As discussed earlier, a floor effect is not a major concern for the Low placement level. Two students in the Low placement level correctly answered all Step 2 items; however, this number is small relative to the total number of students placed in the Low placement level.

In general, the WIDA MODEL Listening placement algorithms were found to do a good job of directing the field test students into a level that appears to be of the appropriate difficulty level. There is a weak indication of a potential ceiling effect for the 6-8 Low placement level. The observed ceiling effect for the Low level is not considered serious, however, since it is unlikely that many students would be affected.

The placement algorithm for the Reading test, also discussed in Chapter 1.3.3 of this report, is as follows: All students complete Reading Step 1; that score is combined with the Writing Quick Score to determine a placement of Low, Mid, or High for Reading Step 2. For Grades 6–8, no students in any placement level answered all Step 2 items incorrectly, and only three students in the Low level, three students in the Mid level, and one student in the High level answered all Step 2 items correctly. The results suggest that there does not seem to be a floor effect at any placement level. There were three students in the Low and Mid level who answered all Step 2 items correctly; however; the number of students at the ceiling for both placement levels is very small. For Grades 9–12, only two students in the Low placement level and three students in the Mid placement level incorrectly answered all items, and only two students in the Low placement level answered all Step 2 items correctly; again, the number of students at the ceiling in the Low placement level is very small.

Overall, these results suggest that the placement algorithms in the WIDA MODEL Reading section generally worked well at directing the field test students into a level that appears to be of the appropriate difficulty level. For both Grades 6-8 and 9-12, there is weak evidence of a ceiling effect for the Low placement level. Since the placement rules were designed such that the majority of students are placed in the Mid level, the observed ceiling effect for the Low level is not as serious since it is unlikely that many students would be affected. However, the observed ceiling effect in Grades 6-8 and floor effect for Grades 9-12 may be problematic.

It is important to note that data used in the placement rule analyses are before the outliers were removed; therefore, these results may have been affected by students whose performances on WIDA MODEL are somehow unexpected.

6.2.4. Evidence from Test Reliability

Because the Overall Composite score and corresponding proficiency level on WIDA MODEL (rather than the domain scores) are used to determine if a student should be placed in ELL services, it is essential to ensure Overall Composite scores are reliable. As mentioned in Chapter 1, the Overall Composite score both on MODEL and MODEL Screener is a composite that weights the individual domains as: 35% Reading + 35% Writing + 15% Listening + 15% Speaking. To obtain the reliability for the Overall Composite scores, stratified alphas were computed based on the Cronbach's alphas of the individual domains and the variance of students' Overall Composite scores for students who had completed all four domains.

Stratified Cronbach's alpha coefficient (Cronbach, Schönemann, & McKie, 1965) is a weighted reliability estimate where the reliability estimate of each domain is weighted by the contribution of each domain score and added into the composite. Specifically, the formula is

$$\alpha_c = 1 - \frac{\sum_{j=1}^{k} w_j \sigma_j^2 (1 - r_j)}{\sigma_c^2}$$

where

k = number of components j
w_j = weight of component j
σ_j^2 = variance of component j
r_j = reliability of component j
σ_c^2 = variance of weighted composite.

This formula requires the estimate of the reliability for each individual domain. For Speaking, Listening, and Reading, Cronbach's alpha was computed using IBM SPSS Statistics version 19 software (2010). For Writing, the Generalizability (G) coefficient from GENOVA (Brennan & Crick, 2003) was used.

Cronbach's alphas for the MODEL Listening and Reading domain were computed based on all students that took the particular placement levels in the field test administration. Speaking and Writing do not have placement levels. Therefore, the number of students used to compute the

Cronbach's alpha of each domain varied depending on the domain and the placement levels in Listening and Reading. (The number of items and students by placement levels of MODEL are reported in Table 19 and Table 20.)

Cronbach's alphas for MODEL Screener by domains were computed based on all students that took each domain of MODEL Screener in the field test. Similar to MODEL, the number of students used to compute the Cronbach's alpha by domains varied by domains. (The number of items and students who took MODEL Screener by domains is reported in Table 46 and Table 47.)

The formula for Cronbach's alpha is

$$\alpha = \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_t^2}\right]$$

where

n = number of items i
σ_i^2 = variance of score on item i
σ_t^2 = variance of total score.

The Generalizability (G) coefficient from GENOVA (Brennan & Crick, 2003) was used to estimate the reliability of the Writing field test scores by applying the Generalizability theory. Generalizability theory (Shavelson & Webb, 1991; Brennan, 2001) was developed to assess reliability of measurement in the presence of multiple sources of error. It provides an analytic procedure to partition total variance in observed scores to two or more sources of variance: in this case, one due to the student and one due to the rater. G theory also provides a coefficient of generalizability based on a particular measurement design that is analogous to the reliability coefficient in Classical Test Theory. The G coefficient is defined as the ratio of the universe score (i.e., the response values of all students on all items) and the observed score variance.

Data from the internal CAL writing scoring meeting (see Chapter 4.2.1) were used in this analysis. MODEL and MODEL Screener Writing consists of two Writing tasks for both Grades 6-8 and Grades 9-12. Both tasks are integrated tasks that measure SIL, LoLA and LoSS. Only the set of student papers that were scored by all five raters were used in the Generalizability analyses. These selected papers are the only ones that were rated by all raters for each task, so they provide the best estimate of variability across raters and papers. A one-facet generalizability

(G) study was first conducted in which a rater facet with five levels was specified in the measurement model. Because it is expected that each student would receive only one rating for a MODEL Writing task in the operational testing, a decision (D) study was then conducted using the same data to obtain the reliability coefficient (G coefficient) based on a single rater. Since field test students took either Task 1 or Task 2 but not both, the G and D studies were conducted separately by task.

6.2.4.1. Reliability of the 6–8 Grade-level Cluster Test

The stratified alphas for the Overall Composite scores for both MODEL and MODEL Screener for the 6-8 grade-level cluster are presented in Table 54. For MODEL, the stratified alpha for the Overall Composite scores are presented by all possible combinations of Listening and Reading placement levels (Low, Mid, and High). Speaking and Writing do not have placement levels. Variances of the weighted Overall Composite scores were computed based only on students who had scores in all four domains. Table 54 shows that the reliability for the Overall Composite scores ranges from .85 to .94 for MODEL, and the reliability for the Overall Composite score is .74 for MODEL Screener.

Table 54
Reliability of Overall Composite for Grades 6–8 for MODEL and MODEL Screener

	Speaking	Listening Placement Level	Reading Placement Level	Writing	Stratified Alpha
MODEL	-	Low	Low	-	.85
		Low	Mid		.86
		Low	High		.91
		Medium	Low		.86
		Medium	Mid		.87
		Medium	High		.93
		High	Low		.88
		High	Mid		.89
		High	High		.94
MODEL Screener					.74

The reliability of MODEL Writing field test scores was estimated using G theory (Shavelson & Webb, 1991; Brennan, 2001). There were 58 calibration papers rated by all five raters for Task 1 and 56 calibration papers were rated by all five raters for Task 2 for Grades 6-8. The results of the G and D study for Grades 6-8 are presented in Table 55 by task. The G coefficients based on one rater, .78 and .76, suggest good reliability associated with the MODEL and MODEL Screener Grades 6-8 Writing scores. The average G coefficient of Task 1 and 2 (.77) is used in computing the stratified alpha for the Overall Composite scores for Grades 6-8.

Table 55
Results of the G and D Study for Writing Grades 6-8

Task	Effect	*df*	Estimated G Study Variance Components	Estimated D Study Variance Components
1	Paper	57	2.59	
	Rater	4	0.05	
	Paper * Rater	228	0.74	
			Universe Score	2.59
			Absolute Error	0.74
			Relative Error	0.79
			Generalizability Coefficient	0.78
2	Paper	55	2.61	
	Rater	4	0.08	
	Paper * Rater	220	0.84	
			Universe Score	2.61
			Absolute Error	0.84
			Relative Error	0.93
			Generalizability Coefficient	0.76

6.2.4.2. Reliability of the 9–12 Grade-level Cluster Test

The stratified alphas for the Overall Composite scores for MODEL and MODEL Screener are presented in Table 47. For MODEL, the stratified alpha for the Overall Composite scores are presented by all possible combinations of Listening and Reading placement levels (Low, Mid, and High). Speaking and Writing do not have placement levels. Table 56 shows that the reliability for the Overall Composite scores ranges from .89 to .94 for MODEL, and the reliability for the Overall Composite score is .84 for MODEL Screener.

Table 56

Reliability of Overall Composite for Grades 9–12 for MODEL and MODEL Screener

	Speaking	Listening Placement Level	Reading Placement Level	Writing	Stratified Alpha
MODEL	-	Low	Low	-	.89
		Low	Mid		.91
		Low	High		.93
		Medium	Low		.90
		Medium	Mid		.91
		Medium	High		.93
		High	Low		.91
		High	Mid		.92
		High	High		.94
MODEL Screener					.84

As with Grades 6–8, the reliability of the Writing field test scores for Grades 9–12 was estimated using only data from the internal CAL writing scoring meeting using G theory. There were 49 calibration papers rated by all five raters for Task 1 and 52 calibration papers were rated by all five raters for Task 2.

The results of the G and D study for Grades 9-12 are presented in Table 57. The G coefficients based on one rater, .84 and .85, suggest good reliability associated with the MODEL and MODEL Screener Grades 9-12 Writing scores. The average G coefficient between Task 1 and 2 (.84) is used in computing the stratified alpha for the Overall Composite scores for Grades 9-12.

Table 57

Results of the G and D Study for Writing Grades 9–12

Task	Effect	*df*	Estimated G Study Variance Components	Estimated D Study Variance Components
1	Paper	48	1.97	
	Rater	4	0.07	
	Paper * Rater	192	0.38	
			Universe Score	1.97
			Absolute Error	0.38
			Relative Error	0.45
			Generalizability Coefficient	0.84
2	Paper	51	3.92	
	Rater	4	0.03	
	Paper * Rater	204	0.67	
			Universe Score	3.92
			Absolute Error	0.67
			Relative Error	0.70
			Generalizability Coefficient	0.85

6.2.5. Evidence from Rater Agreement

Establishing inter-rater agreement is an important step toward producing a reliable and valid assessment of students' writing ability. WIDA MODEL Writing tasks require students to create a response and raters to judge the quality of the student responses, building on their understanding of the construct and the scoring rubric. This is a very complicated process, and many factors (e.g., the ability of the student, the difficulty of the task, the scoring process, the nature of the rating scale, the way in which a rater applies the rating scale, etc.) could affect students' Writing scores. The purpose of rater reliability analysis is to determine whether the rating process and the training materials are working as intended and to examine agreement among raters. Classical inter-rater reliability statistics were computed to provide indications of inter-rater agreement and inter-rater consistency; many-facets Rasch analyses were conducted to examine and understand sources of variability in writing scores.

6.2.5.1. Inter-Rater Agreement

For the Writing scoring, each student's writing paper was scored by at least two different raters during Phase II of the Writing scoring (i.e., the external scoring meeting; see Chapter 4.2.1 for more information). Raters were randomly assigned sets of student papers as the first or second reader. For the inter-rater reliability analysis, all of the paired ratings across all student papers were analyzed together by task. Because different raters scored different sets of student papers and not all of the raters scored all sets, the inter-rater statistics computed do not measure the degree of agreement or disagreement between the same two raters across sets. Rather, they are measures of the degree of agreement or disagreement between the first and second raters across sets.

Inter-rater agreement measures the degree to which two raters assigned the same rating to the same student response. If two raters' scores differed by one raw score point or less, the scores were considered to have *good agreement*. This definition is consistent with the criterion used for qualifying raters and for rescoring writing papers (see Chapter 4.2). If two raters' scores differed by two to three raw score points (using the 18-point scale of 1-, 1, 1+, etc.), the scores were considered to have *sufficient agreement*. If two raters' scores differed by more than three raw score points, the scores were considered to be *discrepant*.

Inter-rater *consistency* measures the degree to which independent raters provide the same relative ordering or ranking of persons or performances being rated. Pearson correlations were computed as indications of the inter-rater consistency between pairs of ratings assigned by raters who scored the same student papers.

The means, standard deviations, the percentages of good agreement (|D|=0–1), sufficient agreement (|D|=2–3), and discrepant ratings (|D|>3), and the Pearson correlation between scores assigned by the first and second rater are reported for Grades 6–8 in Table 58 and for Grades 9–12 in Table 59.

For Grades 6–8, the percentage of good or sufficient agreements for Tasks 1 and 2 was very high (98.44% and 98.02%, respectively). Very small percentages (1.56% and 1.99%, respectively) of pairs of ratings were discrepant. The Pearson correlations between the converted raw score assigned by the first and second rater were .73 and .83, indicating that the ratings were fairly consistent. These results suggest that the rubric and the training materials for Grades 6–8 work as intended.

Table 58

Inter-Rater Agreement for Task 1 and Task 2 for Grades 6–8

			First Read		Second Read		Percent Agreement			
	Number of Papers	Maximum Raw Score	Mean	SD	Mean	SD	Good Agreement |D|=0–1	Sufficient Agreement |D|=2–3	Discrepant |D|>3	Pearson Correlation
Task 1	320	14	6.93	1.83	6.14	1.65	69.06%	29.38%	1.56%	.73**
Task 2	403	16	7.67	2.09	6.97	1.89	74.94%	23.08%	1.99%	.83**

***p*<.01.

For Grades 9–12, the percentage of good or sufficient agreement for Tasks 1 and 2 was very high (97.07% and 98.55%, respectively). Very small percentages (2.92% and 1.45%, respectively) of pairs of ratings were discrepant. The Pearson correlations between the converted raw score assigned by the first and the second rater were .83 for Task 1 and .87 for Task 2, indicating that the ratings were consistent. These results suggest that the rubric and the training materials for Grades 9–12 work as intended.

Table 59
Inter-Rater Agreement for Task 1 and Task 2 for Grades 9–12

			First Read		Second Read		Percent Agreement			
	Number of Papers	Maximum Raw Score	Mean	SD	Mean	SD	Good Agreement |D|=0–1	Sufficient Agreement |D|=2–3	Discrepant |D|>3	Pearson Correlation
Task 1	308	16	5.67	2.33	4.82	2.03	71.75%	25.32%	2.92%	.83**
Task 2	207	13	5.86	2.48	5.12	2.11	72.46%	26.09%	1.45%	.87**

***p*<.01.

6.3. Claim 2: Appropriate Classification of Test Takers According to the WIDA ELP Standards

The development of WIDA MODEL was based on the construct and procedures used for ACCESS, including the ELP Standards and MPIs. As a result of this development process, WIDA MODEL scores are linked to the WIDA ELP levels, allowing for a consistent interpretation of students' abilities.

Claim 2: WIDA MODEL was designed and developed to provide proficiency scores that support appropriate and meaningful interpretations about students' abilities and English language proficiency levels in terms of the WIDA ELP Standards.

6.3.1. Evidence from Test Development Procedures

Content validity (AERA, APA, & NCME, 1999) refers to the adequacy of test items to measure knowledge in a specified content area. Content coverage is used as the first indication of content validity. Content considerations for WIDA MODEL were addressed by the test items and maps, which were based on ACCESS items (see Chapter 2.1). Careful adherence to the test maps guaranteed that the tests would validly measure the construct of English language proficiency as represented in the WIDA ELP Standards and that the tests covered all language domains and proficiency levels of the Standards.

Additional evidence of content validity was provided by a series of qualitative evaluations of WIDA MODEL test content during the WIDA MODEL test development process. The content review (Chapter 2.3) and the pilot testing (Chapter 2.7) were conducted by content experts to help ensure that items contained the appropriate content for a given grade level and proficiency level. The knowledge, expertise, and professional judgments of the experts ultimately ensured that the content of WIDA MODEL formed a legitimate basis upon which to validly derive conclusions about students' English language proficiency.

6.3.2. Evidence from Rasch Analysis

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 that are irrelevant to the focal construct.

Rasch models are confirmatory and assume a strong theoretical grounding for item development. Thus, measures that fit the 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 exhibit multidimensionality. The items that fit are likely to be measuring the single dimension intended by the construct; therefore, misfitting items may contain construct-irrelevant variance. As presented in Chapter 4.1, for the Speaking, Listening, and Reading sections of WIDA MODEL, all items fit the Rasch model well and are productive for measurement according to the infit statistics. These results are a strong indication that WIDA MODEL scores reflect the construct that the tests were designed to measure.

For Writing, inter-rater reliability analyses (Chapter 6.2.5) 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 is clear evidence supporting the claim that the construct-irrelevant variance related to scoring the Writing responses was minimized.

6.3.3. Evidence from Correlations with Other Measures

Because WIDA MODEL and ACCESS were developed using the same WIDA ELP Standards and given previous evidence that ACCESS is a valid measure of students' English language proficiency (Center for Applied Linguistics, 2014), WIDA MODEL scores were expected to correlate with ACCESS scores. In particular, since the decisions about students' placement ELL programs may be partially made based on the MODEL or MODEL Screener Overall Composite scores, it is essential to examine whether these scores correlate with ACCESS Overall Composite scores. As mentioned in Chapter 1, the Overall Composite score on MODEL, MODEL Screener, and ACCESS is a composite that weights the individual domains as follows: 35% Reading + 35% Writing + 15% Listening + 15% Speaking.

A correlation of +1 would indicate a perfect positive linear relationship between variables, and a correlation of -1 would indicate a perfect negative linear relationship. Generally, a correlation of

.90–1.0 is considered very high, .70–.90 is high, .50–.70 is moderate, .30–.50 is low, and 0.0–.30 is little (Hinkle, Wiersma, & Jurs, 1979).

Table 60 shows the Pearson correlations between the MODEL Overall Composite scores and ACCESS Overall Composite scores in Grades 6-8 and Grades 9-12. The moderate to high correlations provide support to the claim that MODEL assesses the construct of English language proficiency in a manner similar to ACCESS.

Table 60
Pearson Correlations: MODEL Field Test Overall Composite Scores and ACCESS Operational Test Overall Composite Scores

	Grade 6-8	Grade 9-12
Pearson Correlation	.82**	.89**
N	398	308

***p*<.01.

Table 61 presents the Pearson correlations between the MODEL Screener Overall Composite scores and the ACCESS Overall Composite scores for Grades 6–8 and 9-12. Correlations are calculated only using data from students who had scores on all four domains in both MODEL Screener and ACCESS. Since only a small number of students took the complete Screener test during the field test, the correlations presented in Table 61 are computed based on a much smaller sample as compared to those reported for MODEL.

Table 61 shows that the Pearson correlations between the MODEL Screener Overall Composite scores and the ACCESS Overall Composite scores are lower than those for MODEL. This may be attributed to the fact that the MODEL Screener test form contains fewer items compared to the complete MODEL test form or to the homogeneity of the student sample that took MODEL Screener. As discussed in Chapter 1 of this report, MODEL Screener provides an overall proficiency level score that can be used for identification and placement in ELL services and for determination of tier placement for ACCESS. Because some precision in measurement of students' English language proficiency is sacrificed as a result of its brevity, however, Overall Composite scores on MODEL Screener should be considered as only one of several indicators in the decision process regarding ELL services.

Table 61

Pearson Correlations: MODEL Screener Field Test Overall Composite Scores and ACCESS Operational Test Overall Composite Scores

	Grade 6-8	Grade 9-12
Pearson Correlation	.65**	.62**
N[13]	63	58

***p*<.01.

Moderately high correlations were expected between the domain scale scores because all of the domain tests in MODEL were administered around the same time on the same students, they all measure closely related constructs, and general English language proficiency should underlie proficiency in the individual domains. Table 62 shows the correlations between the MODEL domain scale scores for Grades 6–8. Overall, correlations are low and range from .38 between the domains of Speaking and Reading to .51 between the domains of Listening and Writing.

Table 62

Pearson Correlations: MODEL Field Test Domain Scale Scores for Grades 6–8 (N = 398)

	Speaking	Listening	Writing	Reading
Speaking	1	.50**	.46**	.38**
Listening		1	.51**	.47**
Writing			1	.49**
Reading				1

***p*<.01.

Table 63 shows the correlations between the MODEL domain scale scores for Grades 9–12. Overall, correlations are low to moderate and range from .42 between the domains of Speaking and Reading to .63 between the domains of Speaking and Writing.

[13] Correlations are calculated only from students who had scores on all four domains in both MODEL and ACCESS. In addition, for Grades 6–8, the ninth graders who took MODEL Screener are excluded from the analyses, because ninth graders cannot take the ACCESS test for Grades 6–8.

Table 63

Pearson Correlations: MODEL Field Test Domain Scale Scores for Grades 9–12(N = 308)

	Speaking	Listening	Writing	Reading
Speaking	1	.57**	.63**	.42**
Listening		1	.56**	.48**
Writing			1	.52**
Reading				1

**p*<.01.

Correlations were not calculated for the domain scores for MODEL Screener because, as previously explained, the Screener was developed to measure students' overall ELP level rather than to assess proficiency in individual domains. As a result of its shorter length, only the Overall Composite score for MODEL Screener was used to conduct analyses.

References

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

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

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

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan R. L., & Crick, J. E. (2003). GENOVA: A generalized analysis of variance system. [Software]. Available from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational and Psychological Measurement*, *25*, 291-312.

DeVellis, R. F. (1991). *Scale development*. Newbury Park, NJ: Sage Publications.

Gottlieb, M., Cranley, M. E., & Oliver, A. (2007). *English language proficiency standards and resource guide (2007 ed)*. Madison, WI: WIDA Consortium.

Gottlieb, M., Cranley, M. E., & Cammilleri, A. (2007). *Understanding the WIDA English language proficiency standards: A resource guide*. Madison, WI: WIDA Consortium.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1979). *Applied statistics for the behavioral sciences*. Chicago: Rand McNally College Publishing Company.

IBM Corp. (2010). IBM SPSS Statistics for Windows (version 19.0). [computer software]. Armonk, NY: IBM Corp.

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.

Kenyon, D. M. (2006). *Development and field test of ACCESS for ELLs®: Technical Report # 1*. Madison, WI: WIDA Consortium.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2010). Facets (version 3.58.0). [Software]. Available from http://www.winsteps.com/facets.htm.

Linacre, J. M. (2011). Winsteps (version 3.70.0.5). [Software]. Available from http://www.winsteps.com/winsteps.htm.

MetriTech, Inc. & the Center for Applied Linguistics. (2011). *WIDA MODEL™ test administration manual*. Madison, WI: Board of Regents of the University of Wisconsin System.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smith, E. V. (2000). Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement, 1*, 303-326.

WIDA Consortium. (2007). *WIDA Consortium English language proficiency standards Grade 6 through Grade 12 (2007 ed).* Madison, WI: Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2011a). Differences among WIDA MODEL, ACCESS for ELLs, and W-APT. In *Comparing WIDA MODEL™, ACCESS for ELLs®, and W-APT™.* Retrieved April 30, 2012 from http://www.wida.us/assessment/comparing.aspx.

WIDA Consortium. (2011b). *WIDA MODEL™ Assessment Training Toolkit CD-ROM* [CD-ROM]. Available from https://www.wceps.org/Store/WIDA/#2

WIDA Consortium. (2011c). *WIDA MODEL™ Test Administration Training DVD* [DVD]. Available from https://www.wceps.org/Store/WIDA/#2

Appendix A: Acknowledgements

We would like to extend our appreciation to the many CAL and WIDA staff members who have supported this work, including:

Melissa Amos, M.A.
Catherine Cameron, M.A.
Mohammed Louguit, Ph.D.
Dorry M. Kenyon, Ph.D.
David MacGregor, Ph.D.
April Maddy, Ed.M.
Katherine Merow, M.A.
Jennifer Renn, Ph.D.
Shauna Sweet, M.S.
Tiffany Yanosky, M.A.
Shu Jing Yen, Ph.D.
Xin Yu, M.A.