World-class Instructional Design and Assessment

Technical Report

Development and Field Test of MODEL™

Grades 1–2 and 3–5

Prepared by:

Tiffany Yanosky, M.A.
Mohammed Louguit, Ph.D.
Stephanie Gibson, M.A.
Shu Jing Yen, Ph.D.
David MacGregor, Ph.D.
Dorry M. Kenyon, Ph.D.
Katharine Merow, M.A.
Catherine Cameron, M.A.

Center for Applied Linguistics

May 31, 2012

Executive Summary

The 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 MODEL™ for grades 1–2 and 3–5. The development and field tests of MODEL™ for kindergarten and grades 6–8 and 9–12 are discussed in technical reports separate from this one.

This report about MODEL™ for grades 1–2 and 3–5 begins by providing background information about the purposes, format, and scores (Chapter 1), describing how the tests were developed (Chapter 2) and field tested (Chapter 3), and presenting technical properties of the field tested items and tasks (Chapter 4). Other chapters explain the linking of MODEL™ to the WIDA English Language Proficiency (ELP) levels and WIDA English Language Proficiency (ELP) Standards (Chapter 5) and the validity and reliability of the test (Chapter 6). The report ends with details about the development and technical properties of MODEL Screener (Chapter 7), which is a shorter and quicker test that serves more limited purposes than does the full-length MODEL™.

Summary Highlights

Background Information (Chapter 1)

MODEL™ is an assessment of ELL students' academic English language proficiency in the four language domains 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) 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, 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 additional time during the school year, and to replace the WIDA-ACCESS Placement Test (W-APT)™.

MODEL™ also has a Screener, which includes all tasks from the Speaking and Writing sections of MODEL™ but fewer items from the Listening and Reading sections. The 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. The Screener, however, cannot be used to determine amount, type, or exiting of ELL services.

In both the full MODEL™ and the 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 the full MODEL™ has multiple-choice items, is administered to individual students, and has placement levels Low, Mid, and High so students take only items that are appropriate for their proficiency level. The Reading section in the full MODEL™ is also multiple choice and has placement levels Low, Mid, and High, but the section is administered to individual students in grades 1–2 and to groups in grades 3–5. The Listening and Reading sections in the Screener have the same format and administration as in the full MODEL™, but they contain fewer items and students are not placed into different levels. The Writing section in both the full MODEL™ and the 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.

After a test administrator completes 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 interpret a student's scale score in terms of the WIDA ELP Standards.

Test Development (Chapter 2)

The MODEL™ tests for grades 1–2 and 3–5 were originally patterned after the ACCESS tests for grades 1–2 and 3–5, as the MODEL™ tests would use folders of items that were retired, or removed, from the ACCESS operational test. However, because not enough retired folders were available, additional folders were selected from the ACCESS field test, were newly created either from an Item Writing Workshop or at the Center for Applied Linguistics, and were adapted from ACCESS by using similar question types, vocabulary, and language structures.

All items underwent a series of reviews—a content review, an international perspectives panel, and a bias and content review—to ensure that items contained the appropriate content for a grade level and proficiency level, that items were appropriate and universal to people of different ethnic backgrounds, and that they did not contain cultural bias or sensitive topics. In addition, five cognitive labs were held to collect information about administration 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 MODEL™ test forms were finalized.

Field Test (Chapter 3)

Field testing for grades 1–2 and 3–5 occurred with 1,264 students in 22 schools in four WIDA states—Alabama, Illinois, Virginia, and Wisconsin—and the District of Columbia from August through October 2009. WIDA hired field test coordinators and field testers to assist with the testing of students. Field testers followed the same procedures, administration, and scoring 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. Rasch analyses revealed that overall these items are productive for measurement and measure what they are intended to measure.

For the Writing sections, sets of writing samples were used to calibrate test development staff at the Center for Applied Linguistics and later consultants who were hired from outside. After all raters had learned how to score the calibration samples, all writing samples were rated. Analyses of these raw scores indicated that all Writing tasks have an appropriate level of difficulty for students.

Linking MODEL™ to WIDA ELP Levels (Chapter 5)

To make the scores on MODEL™ more usable to educators, scores on the test were linked to scores on ACCESS so they can be interpreted in terms of the WIDA English Language Proficiency levels (Level 1 Entering through Level 5 Bridging). A linking study was conducted in order to produce lookup tables, which show for each grade and domain the proficiency level scores that correspond with students' raw scores and scale scores. In the linking study for Listening and Reading, psychometric methods—common item linking and common person linking—and qualitative methods—a bookmarking study—were used to estimate the difficulty measure of items. Those results were applied to the field test data, and resultant growth in student ability was compared to the expected growth. For Writing and Speaking, expert panels qualitatively interpreted performances on MODEL™ to establish a scale.

Validity (Chapter 6)

The validity and assessment use arguments presented in this chapter link students' test performance on MODEL™ to test scores and provides evidence of the interpretation of the test scores. 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' scales scores on MODEL™ were correlated with their scale scores on ACCESS. Although ACCESS had been administered several months earlier, correlations between MODEL™ scale scores and ACCESS scale scores were moderate to high for most domains and the Overall composites.

Analyses of test reliability produced findings that were similar for both grades 1–2 and grades 3–5. For both tests, Cronbach's alphas for Speaking were high. A few Listening and Reading placement levels had lower-than-expected Cronbach's alphas. The reliability of the Writing scores was investigated using a decision (D) study to obtain the reliability coefficient (G-coefficient), and the results indicated that the tasks had good reliability. The stratified alphas for the Overall score—that is, the composite of Speaking, Listening, Reading, and Writing—were above the 0.80 criteria expected from this type of test. The reliability for the Overall score is important because it is used to make decisions about students' English language proficiency and placement in classes.

In addition, for the Writing tasks, inter-rater reliability statistics were computed and indicated high inter-rater agreement and inter-rater consistency. Many-facets Rasch model analyses found that only one rater had a larger-than-expected variability or inconsistency in rating the students' Writing papers. These results suggest that the rubric, scoring procedures, and training materials are sufficient for raters to render reliable Writing scores.

Development and Technical Properties of MODEL Screener (Chapter 7)

This chapter begins by explaining how folders were selected from the full MODEL™ to create the Screener, which contains the entire Speaking and Writing sections from MODEL™ but fewer items in the Listening and Reading sections. Rasch analyses were performed for different scenarios of Mid- and High- level folders for Listening and Reading, and folders that were found to have good fit statistics and to best measure students' English language abilities were selected. In support of the validity of the Screener, students' scale scores on the MODEL™ Listening section and Reading section correlated highly with the students' scale scores on the Screener Listening section and the Screener Reading section, respectively. Because the Screener contains only seven items in Listening and only seven items in Reading, the reliabilities of those sections were slightly lower than the reliabilities of the Listening and Reading sections in the full MODEL™. However, the Overall reliabilities were high, indicating that the Screener can be used to determine if students are eligible for ELL services.

Table of Contents

1. Background

The 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, MODEL™ can be used by educators to identify newly enrolled students as ELLs, to place students in ELL services, or to monitor interim progress.

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. 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: for kindergarten in October 2008, for grades 1–2 and 3–5 in August 2010, and for grades 6–8 and 9–12 in September 2011.

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

1.1. Purposes of 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 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.

1.2. Underlying MODEL™

1.2.1. Alignment with the WIDA ELP Standards

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 English language learners. From its conceptualization to its launch, MODEL™ was planned to be a comprehensive English Language Proficiency exam assessing students' English language proficiency in the five WIDA English Language Proficiency (ELP) Standards:

> Standard 1 - English for **Social and Instructional Language (SIL)** purposes within the school setting;

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

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

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

Standard 5 - information, ideas, and concepts necessary for academic success in the content area of **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, English as a Second Language, 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 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;

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, the MODEL™ test 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 Performance Definitions and the Model Performance Indicators.

1.2.4. Performance Definitions

Performance definitions specify the language that ELLs will process, understand, produce, or use at each of the five defined language proficiency levels. The performance definitions associated with a given proficiency level address desired linguistic attainments for the three levels of language analysis: discourse, sentence, and word/phrase. Figure 1.2.4 associates performance definitions with the proficiency levels laid out in the WIDA ELP Standards.

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 1.2.4: 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, but 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 1.2.5 shows an example of an MPI for the Social and Instructional Language Standard in the MODEL™ Listening section for grades 1–2. This example shows the type of discourse "Following directions" and how elementary school students' comprehension progresses as they move through the levels of English language proficiency 1–5.

	Example Topics	Level 1 Entering	Level 2 Beginning	Level 3 Developing	Level 4 Expanding	Level 5 Bridging
LISTENING	**Following directions**	Follow oral directions according to simple commands using manipulatives or real-life objects (e.g., "Show me your paper.")	Follow oral directions according to complex commands using manipulatives or real-life objects (e.g., "Put the cubes in a row across the paper.")	Follow oral directions by comparing them with visual cues, nonverbal cues or modeling (e.g., "Fold the paper in half. Then place it on your table the long way.")	Follow oral directions without visual or nonverbal support and check with a peer (e.g., "Put your name on the top line of the paper.")	Follow a series of oral directions without support (e.g., "Put your name on the left-hand side of the paper. Then put the date on the right-hand side.")

Figure 1.2.5: A Strand of Model Performance Indicators with an Example Topic
Source: *WIDA Consortium English Language Proficiency Standards PreKindergarten through Grade 5 2007 Edition* (WIDA Consortium, 2007)

1.3. Format of MODEL™

While Chapter 1.2 laid out the organizing principles that underlie MODEL™, Chapter 1.3 describes the practicalities of the test itself.

1.3.1. Grade-level Clusters

MODEL™ has test forms for kindergarten, grades 1–2, grades 3–5, grades 6–8, and grades 9–12. The appropriate test 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 take the test for the previous grade-level cluster if it is the first semester of the school year. For example, as seen in Figure 1.3.1, WIDA recommends that third graders in their first semester take the test form for grades 1–2, while third graders in their second semester, all fourth graders, all fifth graders, and sixth graders in their first semester should take the test form for grades 3–5. WIDA has 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 1.3.1: Appropriate Form of WIDA MODEL™ Based on Grade Level and Semester
Source: *WIDA MODEL™ Test Administration Manual* (MetriTech and CAL, 2010)

1.3.2. Adaptability

Test items and tasks that allow students at proficiency level (PL) 1 or 2 to demonstrate the full extent of their language proficiency may not challenge students at PL 4 or 5. Likewise, items and tasks developed for students at PL 4 or 5 are likely to be far too challenging for students at PL 1 or 2. Items that are too easy for test takers might be boring and lead to inattentiveness, and items that are too difficult for test takers might be frustrating and prevent students from performing their best.

To match the challenge level of tasks to the proficiency level of the test taker, 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 certain parts in Step 2 based on his or her proficiency level. Test administrators determine which placement is appropriate for a student early on rather than assign a placement or make a student work his or her way up. 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 1.3.2 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 because a student at that level is considered able to participate in school without language support. The test and the placement rules are designed so that most students will take the Step 2 Mid placement. Only students who have the very lowest levels of proficiency will take Step 2 Low, and only students who have the highest levels of proficiency will take Step 2 High.

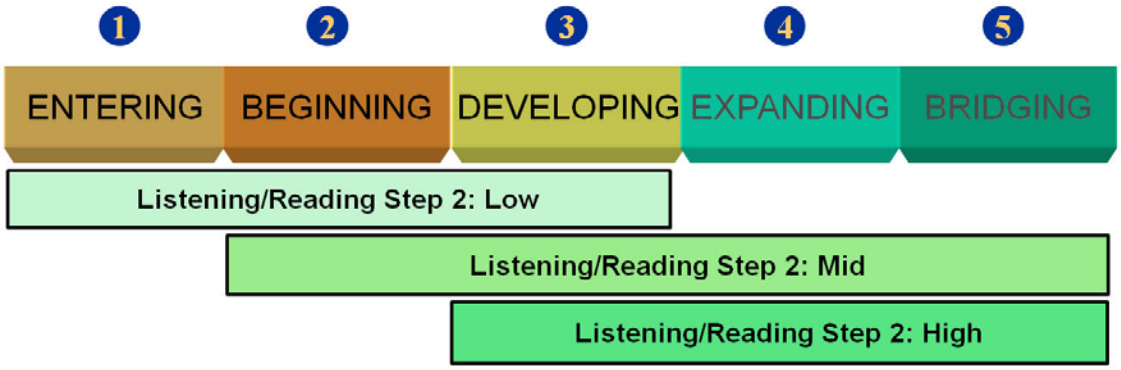

Figure 1.3.2: WIDA ELP Levels and MODEL™ Step 2 Placement Levels
Source: *WIDA MODEL™ Test Administration Manual* (MetriTech and CAL, 2010)

1.3.3. Domain Sections

MODEL™ consists of four domain sections, one for each of Speaking, Listening, Writing, and Reading. Each domain section is organized into "folders," or thematic sets of items or tasks with increasing linguistic demand. Figure 1.3.3 shows the sequence that test administrators follow to administer the domain sections of MODEL™.

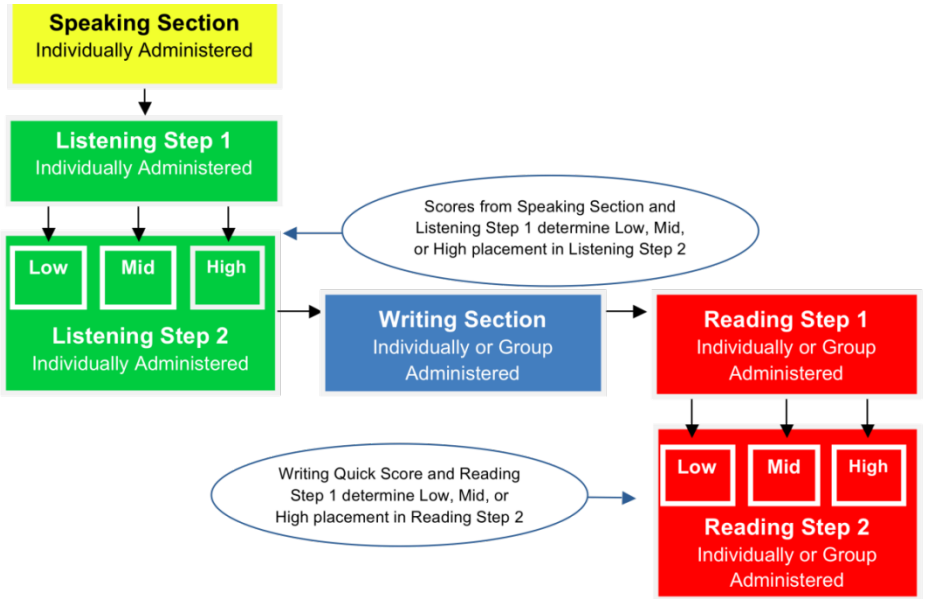

Figure 1.3.3: Administration Sequence of MODEL™ Domains
Source: *WIDA MODEL™ Test Administration Manual* (MetriTech and CAL, 2010)

The test begins with the Speaking section, which is individually administered to students in an interview format. The Speaking section is comprised of two folders, one with three tasks and one with five tasks. These folders address the standards of SIL, LoLA, and LoSS and include tasks targeted at proficiency levels one through five. 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. This section should take less than ten minutes to administer.

The next section, Listening, is also individually administered, and 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 and then Listening Step 1, a folder of four items presented in increasing order of linguistic difficulty. For each item, the student points to his or her answer in the test booklet, and the test administrator records the answer in the Student Response Booklet. Then, a placement of Low, Mid, or High is determined for Step 2 by using results from the Speaking section and from Listening Step 1. Step 2 contains 9–12 items (three or four three-item folders) depending on the form into which the student is placed. The Listening section takes approximately 20 minutes.

In the Writing section of the test, students are presented a task with 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). The Writing section may be group-administered for grade-level cluster 3–5 but not 1–2 because younger English language learners might be overwhelmed in a group and because the administration of the next section, Reading, requires one-on-one administration. The administrator of the Writing section can choose to give a student or students either of two tasks, Writing Task 1 or Writing Task 2. The tasks are about different topics but are meant to elicit the same level of writing. 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. Administration time for a Writing task can last up to 30 minutes. 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 4.2.1.1) and is intended to assist with placement into the appropriate Reading placement level.

The Reading section consists of a series of 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. For grades 1–2, Step 2 contains 15 items (one three-item folder followed by three four-item folders) for the Low level and 12 items (four three-item folders) for the Mid and High levels. The items for the Low level require the student to read single words and then to progress to a task that requires the student to read sentences. For grades 3–5, Step 2 contains 9 items (three three-item folders) for the Low level and 12 items (four three-item folders) for the Mid and High levels. Students in grades 3–5 record their answers by bubbling them in the Student Response Booklet. Students in grades 1–2 point to their answers, and the test administrators write the selection in the Student Response Booklet. Step 2 Low for grades 1–2 is unique in that the students flip over cards with text on them and then match the text to the appropriate picture in the Student Test Booklet. Because of these unique administration requirements, the Reading section may be group-administered for grades 3–5 but must be individually administered for grades 1–2. Administration time for the entire Reading section can last 20–25 minutes.

1.4. The MODEL Screener

The MODEL Screener is a shortened, quicker version of MODEL™ that can be used to determine if a student is eligible for ELL services. The MODEL Screener includes the same Speaking and Writing tasks as MODEL™ but has fewer items in Listening and Reading. The Listening and Reading sections of the Screener consist of the four-item folder in Step 1 and one additional three-item folder for students who meet the requirements to advance to Step 2. Because the main purpose of the Screener is to determine if a student is eligible for ELL

services, a folder from the High placement level is used to ensure that the test is most reliable at the higher levels of proficiency. All items in the MODEL Screener are embedded within the full MODEL™ so that no additional materials are needed. Figure 1.4 highlights the parts of the MODEL™ test that are used for MODEL Screener. (Note that, in this context, A, B, C, and D refer to folders, not tiers.)

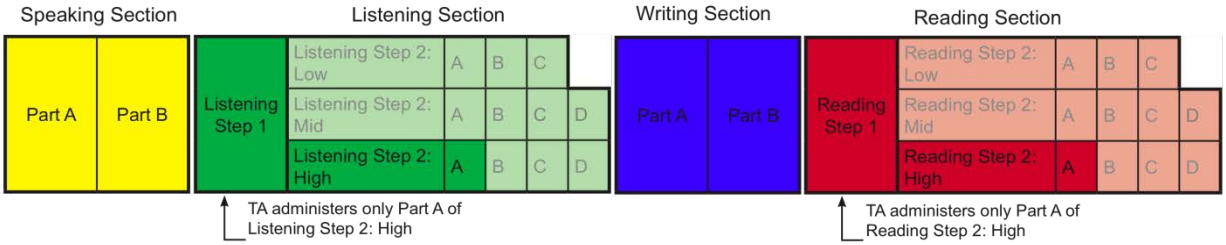

Figure 1.4: Components of the MODEL Screener
Source: *WIDA MODEL™ Test Administration Manual* (MetriTech and CAL, 2010)

Table 1.4 recommends which assessment—MODEL™ or the MODEL Screener—to administer to a student based on the intended purpose. Both MODEL™ and the 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, because the MODEL Screener contains fewer items in Listening and Reading, it is a slightly less reliable test, especially at the lower levels of proficiency, and is, thus, intended to provide only an Overall score and not domain scores. As a result, the Screener cannot provide guidance on the amount and type of ELL services, cannot be used to exit a student from an ELL program, and cannot serve as an interim benchmark assessment. (See Chapter 7 for details about the development and technical properties of MODEL Screener.)

Table 1.4
Differences Between MODEL™ and the 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[1]
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

1.5. Test Scores

MODEL™ scores are reported as both scale scores and proficiency level scores.

[1] Although MODEL Screener does provide an English language proficiency level, this determination is based on less information and therefore is potentially less accurate than the proficiency level provided by MODEL™.

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. Because MODEL™ and ACCESS were developed using the same standards and because a reporting scale had been developed and validated for ACCESS (Kenyon, 2006), MODEL™ scale scores are reported on the same vertical scale as ACCESS through a linking process (see Chapter 5). 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. They consist of a two-digit decimal number. The first digit represents the student's overall language proficiency level based on the student's scale score. The number to the right of the decimal is an indication of the proportion of the range between cut scores that the student's scale score represents. For example, a score of 4.5 indicates that the student's scale score is halfway between the cut for 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 score is based on students' performances in all four domains, it is recommended as the best MODEL™ scale score to use in making educational decisions about students' English language proficiency.

1.6. MODEL™ and ACCESS

Users of MODEL™ who are already familiar with ACCESS or W-APT™ may find it helpful to see the related assessments explicitly compared, as is done in Figure 1.6 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	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, 8 tasks total = up to 10 minutes
Listening	25 minutes	up to 20 minutes	up to 20 minutes
Reading	35 minutes	up to 30 minutes	up to 25 minutes
Writing	60 minutes	15 minutes	up to 30 minutes

Figure 1.6: Differences among WIDA MODEL™, ACCESS for ELLs®, and W-APT Source: *Comparing WIDA MODEL™, ACCESS for ELLs®, and W-APT™* (WIDA Consortium, 2011)

2. Test Development

2.1. Test Maps

During the planning stages of the test development, CAL managers, with stakeholders' input, created a test map for grades 1–2 and grades 3–5 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 1–2 and the test map for grades 3–5 are both shown in Table 2.1, as the test for each grade-level cluster was planned to have the same number of folders for each WIDA ELP Standard and folder tier.

Table 2.1

Test Map for MODEL™ for Grades 1–2 and Grades 3–5

	Listening		Reading		Writing	Speaking
Test Step and Placement Level	**WIDA ELP Standard[2]**	**Folder Tier[3]**	**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	LoLA	C	LoLA	C		
	LoMA	B	LoMA	B		
	LoSC	B	LoSC	B		
	LoSS	C	LoSS	B		
Step 2: High	LoLA	C	LoLA	C		
	LoMA	C	LoMA	C		
	LoSC	C	LoSC	C		
	LoSS	C	LoSS	C		

Folders for MODEL™ were selected from various sources to meet the test map criteria. At the beginning of the test development for MODEL™, there were 6 Listening folders, 4 Reading

[2] As described in Chapter 1.2.1 of this report, the acronyms for the WIDA ELP Standards can be written as 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. For Writing, IT indicates an integrated task that includes SIL, LoLA, and LoSS.

[3] The tier indicates the proficiency level of items in a folder rather than the placement level of the test (Low, Mid, and High). Tier A folders have items at proficiency levels 1 (Entering), 2 (Beginning), and 3 (Developing). Tier B folders have items at proficiency levels 2 (Beginning), 3 (Developing), and 4 (Expanding). Tier C folders have items at proficiency levels 3 (Developing), 4 (Expanding), and 5 (Bridging). Tier B+ folders have items at proficiency levels 2 (Beginning), 3 (Developing), 4 (Expanding), and 5 (Bridging). Note that Speaking and Writing have no tiers, as they are comprised of tasks that are progressively demanding and are thus intended for students of all proficiency levels.

folders, 2 Writing folders, and 1 Speaking folder retired from ACCESS for use in MODEL™ for grades 1–2, and there were 4 Listening, 6 Reading, 3 Writing, and 2 Speaking folders retired from ACCESS for use in MODEL™ for grades 3–5. These retired items would receive new graphics and other necessary revisions.

The remaining folders that were needed to complete the test map for MODEL™ would be items that were new or recently developed or items that were adapted from ACCESS. Entirely new folders were created at the Item Writing Workshop (see Chapter 2.2 below for details), and all of them were written to the specifications used for ACCESS. Other folders for MODEL™ were adapted from existing ACCESS folders and item specifications. Then, using new themes in the same content areas, test developers generated adapted folders with similar question types and similar-frequency vocabulary and language structures. Remaining gaps in the test map were filled by folders that were originally developed for ACCESS series 201 (2009–2010 academic year) but were not needed for ACCESS. These folders had good psychometric properties in their field test, so MODEL™ included these folders.

2.2. Item Writing Workshop

WIDA held an Item Writing Workshop for MODEL™ at the Center for Applied Linguistics from September 26–28, 2008. The goal was to draft items for Reading and Listening, as items for those domains were especially needed.

As seen in Table 2.2A, five item writers developed Reading and Listening folders for grades 1–2. Item writers were assigned to this grade-level cluster based on their experience with teaching and item writing for the grades.

Table 2.2A
Item Writers and Their Affiliations for Grades 1–2

Name	**Affiliated School**
Rachel Howard	Roosevelt Elementary School, Allentown, PA
Margaret Bagnola	Stevenson School, Melrose Park, IL
Paola Lewis	Stevenson School, Melrose Park, IL
Amy Ettner	Danz Elementary School, Green Bay, WI
Patricia Agee-Aguayo	Danz Elementary School, Green Bay, WI

As seen in Table 2.2B, three item writers developed Reading and Listening folders for grades 3–5. One of the schools had a pair of item writers who had been recruited from a paired participant list for an ACCESS item-writing course. The other item writer was an ESL instructor at a university. All item writers who were assigned to this grade-level cluster had prior experience with teaching and item writing for those grades.

Table 2.2B
Item Writers and Their Affiliations for Grades 3–5

Name	Affiliated School
Kelly Boggs	University of Missouri, Columbia, MO
Danielle Kison	Glacier Edge Elementary School, Verona, WI
Erin McGrath	Glacier Edge Elementary School, Verona, WI

Prior to the Item Writing Workshop, participants were given a chart that identified the WIDA ELP Standards, proficiency levels, and language domains for which they would be writing items. Using resources such as textbooks, worksheets, trade books, websites, and colleagues at their schools, item writers began to think of interesting topics from which to build age- and proficiency level-appropriate folders.

Item writers drafted Reading folders on the first day and Listening folders on the second day of the Item Writing Workshop. The following activities were used for developing folders: displaying exemplars, generating a list of descriptors of what students comprehend and produce at each proficiency level, reviewing a Reading Item Review Checklist, walking through a theme folder worksheet and a suggested development order, brainstorming and listing Reading topics for each standard, and drafting and revising the Reading folders.

Item writers generated folders that filled in many of the gaps in the MODEL™ test map. As seen in Table 2.2C, by the end of the Item Writing Workshop, 10 Reading and 6 Listening folders were drafted for grades 1–2. The folders were written to specifications for Tier A, Tier B, and Tier C and addressed four WIDA ELP Standards, LoLA, LoMA, LoSC, and LoSS. Folders for SIL were not developed because retired ACCESS folders could be used instead.

Table 2.2C
Folders Created for Grades 1–2 during the Item Writing Workshop

Domain	WIDA ELP Standard	Folder Tier	Folder Title
Reading	LoMA	A	Buying a Bat
Reading	LoMA	A	Pennies
Reading	LoMA	B	Pencils
Reading	LoMA	B	Market
Reading	LoMA	C	The Birthday Fish
Reading	LoMA	C	At the Carnival
Reading	LoSC	C	Spring Rainfall
Reading	LoSC	C	Frog Life Cycle
Reading	LoLA	C	Baseball Movie
Reading	LoSS	C	Home Sweet Home
Listening	LoMA	A	Positions
Listening	LoLA	A	Getting Ready for School
Listening	LoSC	B	Senses at the Circus
Listening	LoLA	B	Coming Home from School
Listening	LoLA	C	Making Friends
Listening	LoSC	C	Birds

In addition, as seen in Table 2.2D, 8 Reading and 8 Listening folders were drafted for grades 3–5. The folders were written to specifications for Tier A, Tier B, and Tier C folders and addressed four standards, LoLA, LoMA, LoSC, and LoSS. Folders for SIL were not developed because retired ACCESS folders could be used instead.

Table 2.2D

Folders Created for Grades 3–5 during the Item Writing Workshop

Domain	WIDA ELP Standard	Folder Tier	Folder Title
Reading	LoMA	A	Painting Walls
Reading	LoMA	A	Grocery Shopping
Reading	LoMA	B	School Erasers
Reading	LoMA	B	Sharing a Sandwich
Reading	LoMA	C	Volume
Reading	LoMA	C	Rainfall
Reading	LoSC	C	Conductors and Insulators
Reading	LoSC	C	Plant and Animal Cells
Listening	LoLA	A	Missing Card
Listening	LoLA	B	Mystery
Listening	LoSC	B	States of Matter
Listening	LoSC	C	Adaptations
Listening	LoSC	C	Simple Tools
Listening	LoSS	C	Branches of Government
Listening	LoSS	C	Alaska
Listening	LoLA	C	Missing Globe

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 contents. The final list of folders that were used in the test is included in Chapter 2.7 of this report.

2.3. Content Review

CAL held a content review in November 2008 with three content experts: Kelly Boggs, who reviewed the adapted ACCESS folders for grades 3–5, Kate Jerris, who reviewed the Item Writing Workshop folders for grades 3–5, and Rachel Howard, who reviewed all folders for grades 1–2. The purpose of the content review was to ensure that the content is accessible and relevant to the students in the grade level being assessed. The three content experts completed MS Excel workbooks in order to comment on the appropriateness of the folders and to explain changes that should be made. The following questions guided the review:

- Is the content of this folder grade-level appropriate?
- Is the content presented in a way that is grade-level appropriate?
- Is the language used appropriate for the specified proficiency level?
- Are there areas in the domain checklist that you think are not but should be fulfilled in the folder?

The content reviewers made suggestions for the test developers, mostly providing specific examples of ways to revise the text to make it more grade-level appropriate or to resemble text that their students see in the classroom. They also made some suggestions to revise graphics.

2.4. International Perspectives Panel

In January 2009, CAL conducted an international perspectives panel on all MODEL™ folders for grades 1–2 and 3–5 to minimize construct-irrelevant content for international students. As seen in Table 2.4, participants were seven CAL employees representing different nationalities. CAL employees Dorry Kenyon, Jennifer Boryk, and Stephanie Gibson facilitated the discussion. The group examined all folders to identify content that was confusing, inappropriate, unrealistic, inaccurate, unfamiliar, or not universal, particularly for students who may be taking MODEL™ at an international school abroad. The panel decided that text and graphics in 32 folders (10 Listening, 16 Reading, 3 Writing, and 3 Screener[4]) for grades 1–2 and 24 folders (2 Speaking, 8 Listening, 9 Reading, 2 Writing, and 3 Screener) for grades 3–5 needed to have minor revisions.

Table 2.4
Participants of the International Perspectives Panel and Their Country of Origin

Name	Country of Origin
Basra Abdillahi-Chire	Djibouti
Marcos Carvalho	Brazil
Minkyung Lee	Korea
Sarika Mehta	India
Rafael Michelena	Venezuela
Jumana Salem	Jordan
Olesya Warner	Russia

2.5. Bias and Content Review

A bias and content review for folders for grades 1–2 and 3–5 was held at CAL on February 28, 2009. The purpose was for bias and content reviewers to identify potential cultural bias and sensitivity issues so items could be revised as necessary. This bias and content review focused on prejudice and sensitive topics and so differed from the content review held in November 2008, in which grade-level experts ensured that items reflected content that students see in classes (see Chapter 2.3). Qualified applicants were graduate students in a K–12 ESL or TOEFL program and current or former K–12 ESL or TOEFL teachers. As shown in Table 2.5A and Table 2.5B, five consultants participated in the bias and content review for each grade-level cluster. They were affiliated with organizations in Maryland, the District of Columbia, and Virginia. They included 7 whites (6 females and 1 male), 1 Asian (1 female), and 2 blacks (1 female and 1 male).

[4] This Screener refers to the Reading/Writing Screener that was developed but ended up not being included in the field test or operational version of MODEL™. See Chapter 2.6 for further information.

Table 2.5A

Bias and Content Reviewers and Their Affiliations for Grades 1–2

Name	Occupation	Affiliated School
Julie Yoder	ESL	Alexandria City Public Schools, VA
Thomas Kenea	ESL	Baltimore County Public Schools, MD
Cara Rosson	ESL Elementary	Henrico County Public Schools, VA
Praneetha Arthur	ESL	DC Preparatory Academy, DC
Karen Wesley	ESL	H D Cooke Elementary School, DC

Table 2.5B

Bias and Content Reviewers and Their Affiliations for Grades 3–5

Name	Occupation	Affiliated School
Jennifer Heywood	ESL	North Springfield Elementary School, Fairfax County Public Schools, VA
Edith Tress	Non-public Liaison	Baltimore County Public Schools, MD
Zakaria El Homrani	ESL Instructor	Barnard Elementary School, DC
Patricia Bellman	ESOL Specialist	Poplar Tree Elementary School, Fairfax County Public Schools, VA
Pernilla Urps	ESOL Chair	Keene Mill Elementary School, Fairfax County Public Schools, VA

The bias and content reviewers analyzed and discussed the appropriateness of items for a diverse population of elementary English Language Learners. The reviewers followed a checklist to review item content, item bias, layout appearance, graphics content, graphics bias, and color bias. The reviewers also had a list of sensitive topics—such as violent activities, religion, gambling, sexuality, war, poverty, disease, death, and prehistoric times—to avoid so students would not become upset and distracted during the test.

Participants in the bias and content review provided feedback on how to improve 4 Speaking folders, 26 Listening folders, 29 Reading folders, and 10 Writing folders. Test developers applied all of the recommendations to improve the folders, and no folders were rejected outright. The types of changes included simplifying sentences, revising questions to match the model performance indicators (MPIs), revising topics to be more accessible to all students, and modifying graphics to be more grade-level appropriate, more diverse, or more clearly matched to the text.

2.6. Cognitive Labs

The purpose of the cognitive labs was to gather qualitative data on the functioning of MODEL™, including information about the script, items and tasks, and student performances. The cognitive labs addressed questions such as the following:

- How long does each domain section take to administer?
- Does the Literacy Screener[5] accurately place a student in the appropriate level for the Reading and Writing test?
- Does the Literacy Screener test both Reading *and* Writing?
- Are prompts and questions written clearly?
- Are all graphics clear?
- Are all graphics clearly related to the folder?
- Are the tasks engaging?
- Do the tasks allow students to show what they can do?

For each cognitive lab, three researchers—a test administrator, observer 1, and observer 2—collected data on individual test administrations and debriefed afterward. Observer 1 watched the test administration and asked questions to the test administrator, and Observer 2 watched the test administration and asked questions to the student. Both observers took notes on paper. The goal was to test and interview 10 students per form. Notes were later used to make improvements to the tests prior to the field test.

Each cognitive lab is detailed in its own section below, but in summary, cognitive labs 1 and 3 aimed to determine the best way to make a Low, Mid, or High level placement on the domains. Cognitive lab 2 was intended to ensure that items were grade-level appropriate and contained familiar content for students. The main purpose of cognitive lab 4 was to try out new Listening and Reading Screeners. Cognitive lab 5 was primarily used as a confirmatory test of the selected folders and as a collection of writing samples to be used in the final training materials for test administrators.

2.6.1. Cognitive Lab 1: Milwaukee and Whitefish Bay, WI

The first cognitive lab was conducted at Mitchell School in Milwaukee, WI from March 9–12, 2009 and at Richards School and Cumberland School in Whitefish Bay, WI on March 13, 2009. Cognitive interviewers included CAL employees Stephanie Gibson, Jennifer Boryk, Sarika Mehta, and Emily Evans and WIDA staff Robert Kohl, Andrea Cammilleri, Emily Svendsen, and Pakou Vang.

[5] The Literacy Screener was developed but ended up not being included in the field test or operational version of MODEL™. See Chapter 2.6.1 and 2.6.3 for further information.

The main goal of this first cognitive lab was to determine the best way to place students with the Literacy Screener for the Reading and Writing sections. The Literacy Screener was proposed in the original project plan as a component that would allow students to take different levels of items in Reading and Writing than in Listening. The Literacy Screener was a five-minute constructed-response task that occurred after the Listening and Speaking sections but before the Reading and Writing sections. It consisted of three questions of increasing difficulty, which the students were required to read and then to respond to in writing. The first question was a simple WH- question (i.e., why, how) that required a single, high-frequency word as an answer. The second question was supposed to elicit a short phrase or simple sentence, and the last question a more extended response of perhaps 3–4 sentences. After taking the Literacy Screener, students would then be placed into Low, Mid, or High for the Reading and Writing sections of the test.

This first cognitive lab produced the following findings:

- *The version of the Literacy Screener trialed in the cognitive lab did not provide enough information to make an accurate placement decision.* Performance on the Literacy Screener did not distinguish the high-proficiency students from the mid-proficiency students, so the high-proficiency students were not directed to the High level of the Reading and Writing sections.

  Proposed solution: The Literacy Screener was redesigned after cognitive lab 1 to include a combination of constructed-response and selected-response questions. For grades 1–2, the Literacy Screener was a combination of a "Writing Experience," or a completely open-ended writing prompt in which students had the opportunity to write whatever they could, and a folder of three selected-response questions. For grades 3–5, the Literacy Screener was revised to include a combination of two open-ended questions and a selected-response item at proficiency level 5. Test developers hoped that this format would allow for more accurate placement decisions.

- *Group administrations were needed in the Literacy section of the test for Mid and High students.* In the current form of the Literacy section, students would take the Screener, the Reading section, and then the Writing section. Students who were placed into the Low level would take one group of Writing tasks, students placed into the Mid level would take another group of Writing tasks, and students placed into the High level would take a different group of Writing tasks. Each of these groups required the test administrator to read aloud task-specific instructions, meaning students needed to be grouped homogeneously to take the Writing section.

  Proposed solution: To facilitate group administrations and to be more efficient in scheduling them, test developers decided to use only one Integrated (IT) Writing task for Mid and High students instead of different tasks. These tasks also provided the greatest opportunity for students to produce academic writing. Additionally, it was decided that all administrations of the Low levels for Reading and Writing would be conducted one-

on-one, so students at the early stages of language development would not be in a group with higher-proficiency students.

- *Students in grades 1–2 needed more developmentally appropriate items.* In grades 1–2, the students taking the Low level of Reading were frustrated by multi-sentence reading passages. Although the folders were written to item specifications and contained simple sentences and high-frequency vocabulary, they did not allow students who were just beginning to develop literacy skills to demonstrate what they could do. Additionally, one folder about map skills proved to be challenging for students for construct-irrelevant reasons.

  Proposed solution: Test developers decided to model the Low level of Reading for grades 1–2 after sections of the Kindergarten MODEL™, making it an adaptive test using manipulatives. Cards were created because low-proficiency students who are in the first and second grades are still developing beginning literacy skills, and cards that contain a small amount of text are easier and less intimidating to process than are pages full of text. The cards were made of laminated paper with a question number on one side and text on the other side. The revised Low level of Reading consisted of four independent but thematically related parts, each associated with a picture in the test booklet and with a series of three or four Reading items presented on the cards. The student was required to flip over the cards one at a time, read the text, and point to the corresponding picture in the booklet. As the student moved through the pictures, the text on the cards grew increasingly more complex, beginning with single words and ending with single sentences. The Low level of Reading was adaptive in that students would continue to take the next part until they were unable to meet the criteria for moving on. The test developers also reformatted the folder about map skills so it could be more easily directed by the test administrator and used by the student.

- *The Listening placement was not accurate.* The student's Listening placement of Low, Mid, or High was initially determined solely by a student's Speaking score. However, the cognitive lab showed that using only Speaking as a screener did not produce accurate placement decisions for Listening.

  Proposed solution: A decision was made to consolidate the Low, Mid, and High levels on Listening into one test with entrance points and exit points based on student performance. There would no longer be separate Low, Mid, or High forms but instead one form that began with a series of Low folders, then a series of Mid folders, and then a series of High folders. Test administrators would use the Speaking test to establish a floor and then move through the Listening items until they had established a ceiling. A student would also have the opportunity to go two extra folders beyond what had been established as the ceiling, if the student had easily reached that ceiling. It was designed so that students would have the opportunity to be tested on at least four WIDA ELP Standards and four to five folders.

2.6.2. Cognitive Lab 2: Washington, DC

The second cognitive lab was held at CAL on March 31 and April 1, 2009. Three native English-speaking students, who were family members or friends of CAL employees, were tested. Interviewers included CAL employees Jennifer Boryk, Stephanie Gibson, and Sarika Mehta.

The goal of this cognitive lab was to see how native English-speaking students performed on MODEL™ in order to make sure that the test items were grade-level appropriate and accessible to all students. Two first graders took MODEL™ for grades 1–2. One fifth grader took MODEL™ for grades 3–5.

These students were administered all sections of the tests and provided the following feedback:

- *The content, language, and pictures in the tests were familiar and grade-level-appropriate, and only minor changes needed to be made.*

  Proposed solution: Test developers would make the few minor changes to problematic items.

2.6.3. Cognitive Lab 3: Palos Hills, IL

The third cognitive lab was held in Palos Hills, IL at Dorn Primary Center on May 11, 2009 and at Glen Oaks Elementary on May 12 and 13, 2009. CAL had contacted the schools for participation. Participants included CAL employees Jennifer Boryk, Stephanie Gibson, Emily Evans, and Michael Soto and WIDA representatives Carsten Wilmes, Robert Kohl, Andrea Cammilleri, and Jesse Markow.

The cognitive lab included 16 students in grades 1–2 and 12 students in grades 3–5. The goal for this cognitive lab was to try out the revisions that CAL had made to the test forms since the first and second cognitive labs.

The cognitive lab produced the following findings:

- *The idea of having a Literacy Screener needed to be reconsidered*. The cognitive lab indicated that the revised Literacy Screener still did not distinguish the high-proficiency students from other students, as the high-proficiency students were not placed into the High level on the Reading and Writing sections.

  Proposed solution: Test developers decided to add a Reading Screener (soon to be renamed Step 1), a four-item folder that spanned proficiency levels 2–5. The format of the test would also be rearranged so that Writing was administered before Reading, and performance on the Writing section plus performance on the Reading Screener folder would determine placement of Low, Mid, or High in Reading. Furthermore, the Writing section for all levels of students would now be only one task but with two parts: Part A, a

simple pre-writing task to see whether the student should attempt Part B, a more challenging extended writing task. One challenge in organizing the assessment this way would be to ensure that the test administrator's script provides clear guidance for the test administrator and students.

- *Revisions that had been made to the test for grades 1–2 worked well.* The map skills folder that had been redesigned seemed to be more developmentally appropriate for first and second graders. In addition, the cards on the Low level of Reading were found to be grade-level appropriate, to work logistically, and to allow students who were still learning basic literacy a chance to show what they could do. Students who were not ready to read texts of 2–3 sentences could start by reading words and phrases all the way up to more complex sentences.

  Proposed solution: Although the map skills folder was found to work well for students, it was eventually swapped out for a different map skills folder that had appeared on the ACCESS field test and had good item statistics but was overage for ACCESS.

- *The Listening section needed further revisions to enable accurate placement*. The administration of Listening as one untiered test was found to be cumbersome and confusing for test administrators. Additionally, the option for students to continue two folders beyond their ceiling lengthened the test administration time too much.

  Proposed solution: The administration of Speaking and Listening was made analogous to the administration of Writing and Reading in that students' performance on Speaking and their performance on a four-item Listening Screener folder (soon to be renamed Listening Step 1) would determine Listening placement of Low, Mid, or High.

- *Students in grades 3–5 should not fill in bubbles next to their answers in the Reading Student Test Booklet.* The use of the booklets in this way increased the amount of consumable materials and, in effect, the financial cost.

  Proposed solution: Students would fill in the appropriate bubble on a designated answer sheet in their Student Response Booklet instead of selecting their answer choice in the Reading Student Test Booklet.

2.6.4. Cognitive Lab 4: Newport News, VA

The fourth cognitive lab was held at Sedgefield Elementary School in Newport News, VA on June 12, 2009. CAL employees Sarika Mehta, Jennifer Boryk, Stephanie Gibson, and Emily Evans and WIDA employee Carsten Wilmes participated.

For both grades 1–2 and grades 3–5, six students participated in individual administrations of the Speaking, Listening Screener, and Listening tests, and for grades 3–5, two groups of five students each participated in the Writing, Reading Screener, and Reading tests. Students in

grades 1–2 did not participate in this cognitive lab for the Writing, Reading Screener, and Reading tests because their tests are not group-administered.

The main goal of this fourth cognitive lab was to try out the new Listening and Reading Screeners.

Findings included the following:

- *The Listening and Reading Screeners (Step 1) seemed to accurately place students into the Step 2 levels of Low, Mid, and High.* The cognitive interviewers followed the placement instructions of the Screeners and then used the students' ACCESS scores from that school year to confirm the Step 2 placements.

  Proposed solution: Make these Screeners part of the final field test forms.

- *Scripting for small-group administrations of the Reading and Writing sections of the test for grades 3–5 worked well overall but needed some revisions.*

  Proposed solution: Groups should be no larger than five students in order to avoid logistical difficulties. The scripting needed to be edited to seem more natural. Images of students' materials needed to be included in the scripts in order for test administrators to more easily administer the test to students. Finally, in order to ensure that students are taking the most appropriate level of the test form, test administrators need to be given explicit instructions that the High level is really only for students who are ready to exit ELL services and that students should take the Mid level unless they demonstrate otherwise on the Screener folders.

- *The revised answer sheets in MODEL™ for grades 3–5 worked well.* The students taking the test were exposed to the newly introduced answer sheets in the Student Response Booklet. Students did not have trouble with the separate score sheet.

  Proposed solution: Lay out the booklets as such in the final field test forms.

- *One of the Writing tasks for grades 3–5 was not eliciting extended discourse as intended.*

  Proposed solution: Test developers decided to revise this Writing task so it resembled an adapted folder, because adapted folders seemed to work well in the other domains and did not take as much time to write and revise as folders created from scratch.

2.6.5. Cognitive Lab 5: Washington, DC

On July 16, 2009, a fifth and final cognitive lab was conducted at Bancroft Elementary School in Washington, DC. CAL employees Abby Davis, Jennifer Boryk, Sarika Mehta, Stephanie Marcuccio, and Stephanie Gibson participated.

At this point in the test development, MODEL™ was nearly in its field test form. The goals of the fifth cognitive lab were to finalize the scripting, to try out a newly adapted Writing folder for grades 3–5, to collect writing samples on Part A of the Writing tasks to be used in the final training materials, and to get a better sense of how long the test actually would take to administer. Approximately 10 students were tested in grades 1 through 5.

Interviewers learned the following:

- The revised scripting read more naturally and was easier to follow.
- The adapted Writing folder for grades 3–5 elicited the language and the extended response that was expected.
- Several writing samples for Part A were collected to be used in training materials for test administrators. Writing samples for each score point were inserted into the Test Administrator Manuals as examples, and rationales were given for why the samples were scored as such.
- The administration times were found to be approximately 25 minutes total for Speaking and Listening and about one hour total for Writing and Reading.

After this cognitive lab, the test developers had no large concerns and could, therefore, make final preparations for the field test.

2.7. Finalizing the MODEL™ Field Test Forms

Pre-field test key checks were conducted by CAL employees and external consultants in mock test administrations on January 8 and 11, 2009. A final key check was conducted at CAL on July 14, 2009 after the field test. A Project Coordinator oversaw the process, including fails and reconciliation steps, and provided the final list of keys to MetriTech, Inc.

Prior to publication, the 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, Director of the Language Testing Division at CAL.

Tables 2.7A and 2.7B list the final folders that appear on MODEL™ for grades 1–2 and 3–5. 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 in the final test form. As mentioned throughout Chapter 2 of this report, folders were generated from a variety of sources: retired from ACCESS (see Chapter 2.1.), adapted from ACCESS (see Chapter 2.1), taken from the ACCESS field test (see Chapter 2.1), or newly created from the Item Writing Workshop (IWW) (see Chapter 2.2). Folders that appear on MODEL Screener (see Chapter 7) are marked with an asterisk, and folders that appear in multiple placement levels are marked as "repeated."

Table 2.7A

Final List of Folders for Grades 1–2

Domain and Folder Title	Source[6]	WIDA ELP Standard[7]	Folder Tier[8]	Step and Placement Level
Speaking				
Library*	Retired 103	SIL	N/A	N/A
Bears Doing Chores Outside*	Retired 103	LoLA/LoSS	N/A	N/A
Listening				
Art Class*	Retired 103 + one new item	LoLA	B+	Step 1
Around Brookside Community	Retired 103	SIL	A	Step 2: Low
Getting Ready for School	IWW 2008	LoLA	A	Step 2: Low
Shapes at the Park	Adapted	LoMA	A	Step 2: Low
Shapes	Retired 101	LoMA	B	Step 2: Mid
Kitchen	IWW 2008, adapted	LoSC	B	Step 2: Mid
Bingo's Toy (repeated)	Field test 201	LoSS	C	Step 2: Mid
Making Friends (repeated)*	IWW 2008	LoLA	C	Step 2: Mid
Complex Patterns	Retired 101	LoMA	C	Step 2: High
Fish	IWW 2008	LoSC	C	Step 2: High
Bingo's Toy (repeated)	Field test 201	LoSS	C	Step 2: High
Making Friends (repeated)*	IWW 2008	LoLA	C	Step 2: High
Writing				
No Eggs*	New, In house	IT	N/A	N/A
Flying Kites*	New, In house	IT	N/A	N/A
Reading				
Big Balloon*	Retired 103, Adapted	LoLA	B+	Step 1
Nora's Day	Loosely based on retired	LoLA/	A	Step 2: Low
Market	IWW 2008	LoMA	B	Step 2: Mid
Animals	Adapted	LoSC	B	Step 2: Mid
Spring Garden	Adapted	LoSS	B	Step 2: Mid
Eddie and Timmy (repeated)	Adapted	LoLA	C	Step 2: Mid
At the Fun Fair	IWW 2008	LoMA	C	Step 2: High
Lady Bug Life Cycle	IWW 2008	LoSC	C	Step 2: High
Musical Instruments of the World*	IWW 2008	LoSS	C	Step 2: High
Eddie and Timmy (repeated)	Adapted	LoLA	C	Step 2: High

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

[7] As described in Chapter 1.2.1, the acronyms for the WIDA ELP Standards can be written as 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. For Writing, IT indicates an integrated task that includes SIL, LoLA, and LoSS.

[8] The folder tier correlates to the proficiency level of items in a folder rather than to the placement level of the test (Low, Mid, and High). Tier A folders have items at proficiency levels 1, 2, and 3. Tier B folders have items at proficiency levels 2, 3, and 4. Tier C folders have items at proficiency levels 3, 4, and 5. Tier B+ folders include items at proficiency levels 2, 3, 4, and 5. Placement levels Low, Mid, and High were originally named after ACCESS tiers A, B, and C with B+ indicating Step 1.

Table 2.7B
Final List of Folders for Grades 3–5

Domain and Folder Title	Source	WIDA ELP Standard	Folder Tier	Step and Placement Level
Speaking				
Tina Yang Lunch*	Retired 103	SIL	N/A	N/A
Ernesto's Classroom*	Retired 103	LoLA/LoSS	N/A	N/A
Listening				
Mystery*	IWW 2008	LoLA	B+	Step 1
Following Instructions	Retired 101	SIL	A	Step 2: Low
Missing Card	IWW 2008	LoLA	A	Step 2: Low
School Lunch	Adapted	LoMA	A	Step 2: Low
School Supply Store	Field test 201	LoMA	B	Step 2: Mid
States of Matter	Adapted	LoSC	B	Step 2: Mid
Oregon Trail	Adapted	LoSS	C	Step 2: Mid
Missing Globe (repeated)*	IWW 2008	LoLA	C	Step 2: Mid
Trees	New, In house	LoMA	C	Step 2: High
Adaptations	IWW 2008	LoSC	C	Step 2: High
Alaska	IWW 2008	LoSS	C	Step 2: High
Missing Globe (repeated)*	IWW 2008	LoLA	C	Step 2: High
Writing				
Family Activities*	New, In house	IT	N/A	N/A
Lion and Mouse*	Retired 103	IT	N/A	N/A
Reading				
Canoe Adventure*	Retired 200	LoLA	B+	Step 1
New Book	Adapted	LoLA	A	Step 2: Low
Painted Walls	IWW 2008	LoMA	A	Step 2: Low
Rocks	Retired 101	LoSC	A	Step 2: Low
School Erasers	IWW 2008	LoMA	B	Step 2: Mid
Observing Birds	Retired 101	LoSC	B	Step 2: Mid
The Kingdom of Mali	Adapted	LoSS	B	Step 2: Mid
Nurses (repeated)*	Adapted	LoLA	C	Step 2: Mid
Learning about Weather	IWW 2008	LoMA	C	Step 2: High
Plants and Animal Cells	IWW 2008	LoSC	C	Step 2: High
Ancient Civilizations	Adapted	LoSS	C	Step 2: High
Nurses (repeated)*	Adapted	LoLA	C	Step 2: High

3. Field Test

3.1. Design of the Field Test

The field test for MODEL™ was conducted in the fall of 2009. The purpose of the field test was to collect data on items and tasks in order to examine their psychometric properties, to link MODEL™ field test scores to ACCESS operational scores, and to analyze the validity and reliability of the tests. Schools that participated in the MODEL™ field test included only schools that were in the WIDA Consortium, as ACCESS is administered only in member states, and

schools that were not already participating in the ACCESS field test, as conducting two simultaneous field tests at a school would be difficult. Test developers planned to administer MODEL™ to students a short period of time following the operational ACCESS administration, but due to challenges of recruiting schools, the schools that participated in the field test took ACCESS up to several months prior to the field test.

To have sufficient data to conduct psychometric analyses, test developers aimed to assess 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 small group of test administrators, called field test coordinators (FTCs), who lived locally to the participating schools. Then, these FTCs would train their own groups of test administrators, called field testers (FTs), who were also local to the participating schools, to give the field test to students.

3.2. Participating Schools

MODEL™ was field tested in four WIDA states—Alabama, Illinois, Virginia, and Wisconsin—and the District of Columbia from August through October 2009.

Table 3.2A lists the schools that participated in the field test. Twenty-two schools from eight school districts participated.

Table 3.2A
Schools that Participated in the 2009 MODEL™ Field Test

Dates	District	School
9/1–9/2, 9/8	Palos Hills, IL	Glen Oaks Elementary School
8/31–9/1, 9/9	Palos Hills, IL	Oak Ridge Elementary School
9/10–9/11	Palos Hills, IL	Conrady Junior High
9/14–9/18	Green Bay, WI	Jefferson Elementary School
9/14–10/8	Green Bay, WI	Danz Elementary School
9/20–9/29	Green Bay, WI	Howe Elementary School
9/30–10/12	Green Bay, WI	Ft. Howard Elementary School
9/30–10/7	Green Bay, WI	Edison Middle School
10/6–10/7	Green Bay, WI	Chappel Elementary School
10/9	Green Bay, WI	MacArthur Elementary School
9/29–10/2	Diamond Lake, IL	Diamond Lake Elementary School
10/6–10/7	Shelby County, AL	Montevallo Elementary School
10/8–10/16	Shelby County, AL	Thompson Intermediate School
10/12–10/23	Shelby County, AL	Creek View Elementary School
10/19–10/23	Shelby County, AL	Thompson 6th Grade Center
10/5–10/23	Elgin, IL	Ontarioville School
10/8–10/9	Newport News, VA	Epes Elementary School
10/12–10/13	Newport News, VA	Greenwood Elementary School
10/13–10/14	Norfolk, VA	Little Creek Elementary School
10/19–10/20	Newport News, VA	McIntosh Elementary School
10/20–10/21	Newport News, VA	Lee Hall Elementary School
10/8–10/23	District of Columbia	Brightwood Elementary School

The field test was structured so that each geographic region had between one and three Field Test Coordinators (FTCs). These FTCs were trained on test administration procedures by CAL and then returned to their regions to organize their own team of field testers (FTs). The FTCs had personal and professional connections in the regions, so they were assigned the role of recruiting and training their local teams of FTs.

As seen in Table 3.2B, five FTCs and CAL were responsible for training 22 FTs from Illinois, Wisconsin, Alabama, Georgia, Maryland, Virginia, and the District of Columbia. FTCs and FTs were required to submit a teacher certificate that was awarded in the past six months or to pass a criminal background check conducted by an outside company, Sentry Link LLC. In addition, FTCs and FTs had to pass Human Subjects training through the National Institutes of Health (NIH) Office of Extramural Research. CAL approved all qualified FTCs and FTs.

Table 3.2B
Field Test Coordinators (FTCs) and Field Testers (FTs)

Field Test Coordinator	Field Tester	State
Hillary Marzec	James Marzec	IL
	Jocelyn Marzec	IL
	Michael Marzec	IL
	Michael Roszyk	IL
Janet Postier, Gloria Peterson, and Patricia Griedl	Fred Hayworth	WI
	Jill Vandehey	WI
	Jennifer VanHandel	WI
	Cindy Sweeney	WI
	Jennifer Hanson	WI
	Laury Krause	WI
	Ann Schob	WI
Pat Allison	Angela Adams	AL
	Aimee Domingue	AL
	Rhonda Gregg	AL
	Linda Iza	GA
CAL	Cheryl (Regina) Barber	MD
	Dahlia Hamza Constantine	VA
	Tina Tillman	VA
	Robert Smith Jr.	VA
	Betty H. Tillman	VA
	Elena Kitzantides	VA
	Monique Markham	DC

MODEL™ was field tested from August through October 2009 on a total of 1,264 public school students in grades 1–5. Table 3.2C shows the breakdowns of examinees by grade-level cluster, domain, and placement level of test form for Listening and Reading. The number of examinees for Speaking and Writing are shown in Table 3.2D, as these sections of the test do not have placement levels. Note that the numbers of students vary by domain. In some instances, students were absent when different parts of the test were administered on different days or different parts

of the school day. These absent students are shown in the Missing columns. In addition, one student who took the Reading section for grades 1–2 took only Step 1 and therefore does not have a level of test form, so he or she is shown in the N/A column.

Table 3.2C

Distribution of Field Test Examinees by Level of Test Form for Listening and Reading

		Number of Examinees per Level of Test Form					
Grade-Level Cluster	**Domain**	**Low**	**Mid**	**High**	**N/A**	**Missing**	**Total**
1–2	Listening	68	293	213	0	2	576
1–2	Reading	120	391	60	1	4	576
3–5	Listening	89	278	315	0	6	688
3–5	Reading	109	472	93	0	14	688

Table 3.2D

Number of Field Test Examinees for Speaking and Writing

Grade-level Cluster	Domain	Number of Examinees	Number Missing	Total
1–2	Speaking	575	1	576
1–2	Writing	517	59	576
3–5	Speaking	684	4	688
3–5	Writing	656	32	688

Table 3.2E shows the demographic characteristics of the students by grade-level cluster. To obtain this information, the students' MetriTech IDs from MODEL™ were matched to MetriTech IDs from the operational ACCESS series 200 (2008–2009 academic year) test. Not all MetriTech IDs in MODEL™ were included in ACCESS data, but the available demographic data still provide useful information about the sample. The population was roughly split by gender, and a majority of students were Hispanic and from Wisconsin and Illinois.

Table 3.2E
Demographics for the Field Test Sample by Grade-level Cluster

	Grade-level Cluster 1–2		Grade-level Cluster 3–5	
	N	**P**	**N**	**P**
Sex				
Female	257	44.6%	313	45.5%
Male	284	49.3%	357	51.9%
Missing	35	6.1%	18	2.6%
Race/Ethnicity				
Asian/Pacific Islander	52	9.0%	77	11.2%
Black, non-Hispanic	5	0.9%	12	1.7%
Hispanic	407	70.7%	524	76.2%
American Indian/Alaskan Native	1	0.2%	3	0.4%
Multi-racial/Other	0	0.0%	1	0.1%
White, non-Hispanic	76	13.2%	53	7.7%
Missing	35	6.1%	18	2.6%
State				
Alabama	50	8.7%	121	17.6%
District of Columbia	20	3.5%	31	4.5%
Illinois	221	38.4%	183	26.6%
Virginia	12	2.1%	31	4.5%
Wisconsin	238	41.3%	304	44.2%
Missing	35	6.1%	18	2.6%

3.3. Administration of the Field Test

Prior to the start of the field test, schools submitted the names and some demographic information of students who would participate, and MetriTech printed labels for each student's test booklet.

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 Booklet contained the sections for the test administrators to record the students' answers, the writing prompts, and, for grades 3–5, space for the students to bubble in their Listening and Reading answers. A test administration manual was prepared for all test administrators. Scripts, which contained the passages to be read to the students for the Speaking and Listening sections as well as instructions for the Writing and Reading sections, were also included. Materials were mailed to schools from CAL.

Reference materials, such as test administration manuals and supporting MS PowerPoint presentations, were mailed to Field Test Coordinators. Two weeks later, the FTCs were trained by the MODEL™ test development team and an administrator training team in CAL offices for two days. Upon completion of the FTC training, the FTCs returned to their regions and recruited a team of field testers (FTs), who were trained on test administration procedures. Teams were assembled and trained prior to the first day of scheduled field testing. After schools received the test materials, the FTCs took inventory of the materials and prelabeled all student booklets.

The planned field test and operational test administration for MODEL™ were largely based on the test administration for ACCESS. For example, for general room setup and testing procedures, the test administrator ensured that desks in the testing room were arranged so students could see and hear the test administrator, 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 MODEL™ field test was administered by the FTs in the following sequence: Speaking, Listening, Writing, and Reading. (See Chapter 1.3.3 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. For students in grades 3–5, it was highly recommended to give the assessment 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. Upon completion of the field testing, FTCs mailed back 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 they 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 his or her answers to multiple-choice items, the test administrator recorded the answers in the Student Response Booklet and marked the items as Correct or Incorrect based on the answer keys. At the end of the testing session, the test administrator computed the total number of correct answers to determine the raw score.

3.4.3. Scoring the Writing Section

During the test administration, the test administrator used the scoring criteria in the Student Response Booklet to assign a Quick Score of Low, Mid, or High to students' writing responses. A response received a Quick Score of Low if the student did 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 comprehensible. 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, 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 and consultants according to the WIDA Consortium's Writing Rubric (see Chapter 4.2).

3.4.4. Scoring the Reading Section

The Reading section on MODEL™ had multiple-choice questions that were scored as Correct or Incorrect. Reading Step 1 was scored during test administration because these results helped to determine placement of Low, Mid, or High for Reading Step 2. In Reading for grades 1–2, the student pointed to his or her answer, and the test administrator recorded the answer in the Student Response Booklet. In Reading for grades 3–5, the student filled in the bubble for his or her answer in the Student Response Booklet. After the testing session, the test administrator marked each item in Reading Step 2 as Correct or Incorrect, using the keys located in the Test Administrator's Manual. Then the test administrator recorded the total number of correct answers or total raw score as well as the placement in the Student Response Booklet and transferred all raw scores to the student's Summary Score Sheet.

4. Field Test Results

Field test data were scored and analyzed to examine the psychometric quality of the MODEL™ tasks and items. For the Speaking, Listening, and Reading sections, Rasch analyses were used to examine how the tasks and items function. For the Writing section, many-facet Rasch analyses were used to analyze the students' responses and the difficulty levels of the tasks.

4.1. Results for Speaking, Listening, and Reading

Before the field test data for MODEL™ were finalized, multiple stages of data cleaning, processing, and quality checks were conducted. For the Speaking, Listening, and Reading sections, CAL employed data entry specialists for several days to copy answers from the Student Response Booklets to scannable forms that could easily capture the students' data electronically. The answers from each student's booklet were copied to scannable forms twice, that is, by two separate data entry specialists. Next, a CAL research assistant 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 MS Excel, where data were cleaned by comparing the data from each student's two scannable forms. Discrepancies between each student's two scannings were manually corrected as necessary. In addition, corrections were made if a test administrator had made mistakes in filling in the booklets (e.g., blank or multiple answers) or if a MetriTech student ID number was missing. Furthermore, for quality control, two CAL employees compared answers in a sample of original answer booklets to the answers in the Excel workbook.

After the field test data were finalized, student's original responses for all domains except for Writing were converted into "1" and "0" for ease of psychometric analysis. For Speaking items, "Meets" was coded as "1" and "Approaches" as "0". For Listening and Reading items, "Correct" was coded as "1" and "Incorrect" as "0."

4.1.1. Rasch Analyses for 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

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

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., 50 percent probability of getting it right and 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 measure 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 our measurement model may be considered, psychometrically speaking, to be very strong measures. Various Rasch item statistics for MODEL™ were computed and analyzed to examine whether items are considered strong measures.

In the tables in this chapter of the report, the first column shows the ITEM NAME. Each part of each item name provides some detail about the item. For example, for the Speaking test for grades 1–2, the name of the first item "3600_SIp1g12Library_Part_A_T1" indicates the item number "3600" from the ACCESS item database from which the item was taken, the WIDA ELP Standard "SI" for Social and Instructional Language, the proficiency level "p1" for Entering, the grade-level cluster "12" for grades 1–2, the folder title "Library", the first folder "Part_A" of two folders of the Speaking test, and the number of task "T1" in the folder. In contrast to Speaking, the names of items for Listening and Reading often begin with the letter of the folder tier, A, B, or C.

The second column shows the SCORE, which is the number of examinees who answered the item correctly.

The third column shows the COUNT, the total number of examinees in the analysis for that item. This count of examinees varies from folder to folder for Listening and Reading because different students took different folders depending on their placement level. In addition, a few students might not have been administered an item or have been able to provide an answer.

The fourth column shows the P-VALUE of the item, that is, the percentage of examinees 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 and positive measure indicates a difficult item, and a large and negative measure indicates an easy item. These measures represent the final estimates for each item after anchoring them to their values as based on Common Items or Bookmark procedures (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 consistent, unusual performance in relation to the item's difficulty measure. They measure the degree to which examinees' responses to items deviate from expected responses. Both statistics have an expectation of 1.00. 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 examinees) with a higher-than-desirable fit statistic are referred to as misfitting, while items (or examinees) with a lower-than-desirable fit statistic are referred to as overfitting. Infit is weighted and is less sensitive to outliers than is outfit. 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, or high-level or low-level, proficiency do not perform as expected.

4.1.1.1. Rasch Analyses for the 1–2 Grade-level Cluster

Results of the Rasch item analysis for grades 1–2 Speaking are reported in Table 4.1.1.1A. The p-value and measure columns show that within the two folders—Part A and Part B—of the assessment, each Speaking task was more difficult than the one before it, and the first folder had slightly more students answer tasks correctly than did the more difficult second folder. Such patterns are consistent with the adaptive test design.

Table 4.1.1.1B summarizes the infit and outfit findings. These infit and outfit mean square statistics are indicators for how well the data fits the Rasch measurement model. All 8 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. Only one item, the third item, has an outfit mean square statistic greater than 2.0, indicating potentially distorting or degrading measurement. Further examinations of this particular item revealed that some students had unusual response patterns, for example, getting all of the items in the first folder correct but all of the items in the second folder wrong. The large outfit statistic is spurious in this case.

Table 4.1.1.1A

Rasch Item Analysis: Grades 1–2 Speaking

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.3600_SIp1g12Library_Part_A_T1	571	575	0.99	-9.70	0.89	0.03
2.3601_SIp2g12Library_Part_A_T2	542	575	0.94	-3.80	0.61	0.13
3.3602_SIp3g12Library_Part_A_T3	482	575	0.84	0.41	1.27	9.90
4.2994_LSp1g12CleaningBears_Part_B_T1	564	575	0.98	-7.26	1.33	1.48
5.2988_LSp2g12CleaningBears_Part_B_T2	541	575	0.94	-3.69	0.87	0.23
6.2979_LSp3g12CleaningBears_Part_B_T3	457	575	0.79	1.87	0.60	0.24
7.2973_LSp4g12CleaningBears_Part_B_T4	347	575	0.60	7.60	0.50	0.24
8.3035_LSp5g12CleaningBears_Part_B_T5	276	575	0.48	14.56	0.98	0.23

Table 4.1.1.1B

Distribution of Mean-Square Fit Statistics: Grades 1–2 Speaking

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0	N = 0	N = 1
“distorting or degrading measurement”	% = 0%	% = 12.5%
> 1.5–2.0	N = 0	N = 0
“unproductive but not degrading”	% = 0%	% = 0%
0.5–1.5	N = 8	N = 1
“productive for measurement”	% = 100%	% = 12.5%
< 0.5	N = 0	N = 6
“less productive but not degrading”	% = 0%	% = 75%
Total	N = 8	N = 8
	% = 100%	% = 100%

Table 4.1.1.1C presents the results of the Rasch analyses on the 31 Listening items for the 1–2 grade-level cluster, and Table 4.1.1.1D summarizes the findings. The first four items are from Step 1, and the latter items are for the Low, Mid, and High placement levels of Step 2. All items fit the Rasch model well and are productive for measurement.

Table 4.1.1.1C

Rasch Item Analysis: Grades 1–2 Listening

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.A_222_LAp2g12_ArtClass_T1	391	574	0.68	-0.68	0.96	0.92
2.A_223_LAp3g12_ArtClass_T2	434	574	0.76	-1.22	0.94	0.89
3.A_224_LAp4g12_ArtClass_T3	354	574	0.62	-0.32	0.96	0.95
4.A_LAp5g12_ArtClass_T4	391	574	0.68	-0.68	1.07	1.12
5.A_218_SIp1g12_BrooksideCommunity_Part_A_T1	64	68	0.94	-4.81	1.09	0.86
6.A_219_SIp2g12_BrooksideCommunity_Part_A_T2	51	68	0.75	-3.30	1.03	0.98
7.A_220_SIp3g12_BrooksideCommunity_Part_A_T3	52	68	0.76	-2.84	0.82	0.71
8.A_MAp1g12_ShapesAtThePark_Part_B_T4	67	68	0.99	-7.50	1.00	1.00
9.A_MAp2g12_ShapesAtThePark_Part_B_T5	53	68	0.78	-3.12	0.82	0.64
10.A_MAp3g12_ShapesAtThePark_Part_B_T6	13	68	0.19	0.22	1.13	1.48
11.A_LAp1g12_GettingReadyforSchool_T7	61	68	0.90	-4.26	1.10	0.71
12.A_LAp2g12_GettingReadyforSchool_T8	52	68	0.76	-3.02	0.80	0.72
13.A_LAp3g12_GettingReadyforSchool_T9	44	68	0.65	-2.30	0.95	0.84
14.B_2118_MAp2g12_Shapes_Part_A_T1	261	293	0.89	-2.39	0.96	0.85
15.B_MAp3g12_Shapes_Part_A_T2	274	293	0.94	-2.85	0.97	0.86
16.B_MAp4g12_Shapes_Part_A_T3	232	293	0.79	-1.42	1.00	1.04
17.B_SCp2g12_Kitchen_Part_B_T4	150	293	0.51	0.02	1.11	1.12
18.B_SCp3g12_Kitchen_Part_B_T5	242	293	0.83	-1.67	1.05	1.06
19.B_SCp4g12_Kitchen_Part_B_T6	284	293	0.97	-3.65	1.01	1.03
20.B_LAp3g12_MakingFriends_Part_C_T7	366	506	0.72	-0.65	1.01	1.05
21.B_LAp4g12_MakingFriends_Part_C_T8	387	506	0.76	-0.90	0.91	0.82
22.B_LAp5g12_MakingFriends_Part_C_T9	329	506	0.65	-0.25	1.05	1.12
23.B_4258_SSp1g12_Bingo'sToy_Part_D_T10	291	506	0.58	-0.15	0.96	0.96
24.B_4259_SSp2g12_Bingo'sToy_Part_D_T11	271	505	0.54	0.48	1.06	1.06
25.B_4260_SSp3g12_Bingo'sToy_Part_D_T12	296	506	0.58	0.29	1.07	1.12
26.C_MAp3g12_ComplexPatterns_Part_A_T1	201	213	0.94	-2.05	0.97	0.69
27.C_2121_MAp4g12_ComplexPatterns_Part_A_T2	203	213	0.95	-2.25	1.00	1.00
28.C_2124_MAp5g12_ComplexPatterns_Part_A_T3	146	213	0.69	0.17	1.02	1.01
29.C_SCp3g12_Fish_Part_B_T4	136	213	0.64	0.41	0.94	0.94
30.C_SCp4g12_Fish_Part_B_T5	189	213	0.89	-1.25	1.01	0.87
31.C_SCp5g12_Fish_Part_B_T6	186	213	0.87	-1.11	1.00	1.02

Table 4.1.1.1.D

Distribution of Mean-Square Fit Statistics: Grades 1–2 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 = 31 % = 100%	N = 31 % = 100%
< 0.5 "less productive but not degrading"	N = 0 % = 0%	N = 0 % = 0%
Total	N = 31 % = 100%	N = 31 % = 100%

Table 4.1.1.1E presents the results of the Rasch analyses of the 40 Reading items for the 1–2 grade-level cluster, and Table 4.1.1.1F provides the summary of fit statistics. The first four items are from Step 1, and the latter items are for the Low, Mid, and High placement levels of Step 2. According to the infit, all items fit the Rasch model well and are productive for measurement. For the outfit statistic, 27 items are productive for measurement. There were 8 items that have very low outfit, overfitting the Rasch model but not being degrading. That is, they provide redundant measurement information rather than additional measurement information. Furthermore, 4 items have outfit that is less productive but not degrading. One item has an outfit mean square statistic that is greater than 2.0, indicating potentially distorting or degrading measurement.

Table 4.1.1.1E

Rasch Item Analysis: Grades 1–2 Reading

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.2895_LAp3g12_BigBalloons_T1	542	571	0.95	-2.76	1.13	1.58
2.2896_LAp4g12_BigBalloons_T2	449	570	0.79	-1.66	1.43	1.51
3.2897_LAp4g12_BigBalloons_T3	393	567	0.69	-1.25	1.15	1.98
4.2898_LAp5g12_BigBalloons_T4	395	567	0.70	-1.27	1.13	2.34
5.Part_A1_bed	97	120	0.81	-2.83	0.98	1.62
6.Part_A2_cup	104	120	0.87	-2.99	0.99	0.86
7.Part_A3_doll	82	120	0.68	-2.51	1.14	0.91
8.Part_B1_he runs	75	120	0.63	-2.36	0.76	0.56
9.Part_B2_toy car	78	120	0.65	-2.43	0.68	0.46
10.Part_B3_open bag	74	120	0.62	-2.34	0.84	0.73
11.Part_B4_they sit	76	120	0.63	-2.38	0.77	0.64
12.Part_C1_the flower is tall	56	120	0.47	-1.93	0.51	0.36
13.Part_C2_the hat is white	31	120	0.26	-1.23	0.95	0.64
14.Part_C3_they ride bikes	55	120	0.46	-1.91	0.66	0.57
15.Part_C4_cat is under tree	58	120	0.48	-1.98	0.64	0.47
16.Part_D1_chiks eat seeds	39	120	0.33	-1.48	0.56	0.36
17.Part_D2_animal stands	32	120	0.27	-1.26	0.78	0.47
18.Part_D3_three ducks swim	44	120	0.37	-1.62	0.57	0.39
19.Part_D4_farmer watches rooster	26	120	0.22	-1.05	0.62	0.33
20.2854_SSp2g12_SpringGarden_Part_A1	368	397	0.93	-1.94	0.97	0.99
21.2855_SSp2g12_SpringGarden_Part_A2	377	396	0.95	-2.14	0.91	0.68
22.2856_SSp2g12_SpringGarden_Part_A3	339	396	0.86	-1.56	0.89	0.70
23.MAp2g12_Market_Part_B4	345	392	0.88	-1.65	1.01	1.23
24.MAp3g12_Market_Part_B5	210	389	0.54	-0.49	1.22	1.32
25.MAp4g12_Market_Part_B6	182	389	0.47	-0.27	1.07	1.06
26.SCp2g12_Animals_Part_C7	317	387	0.82	-1.38	0.92	0.82
27.SCp3g12_Animals_Part_C8	288	387	0.74	-1.12	0.94	0.85
28.SCp4g12_Animals_Part_C9	205	387	0.53	-0.46	1.10	1.19
29.LAp3g12_EddieandTimmy_Part_D10	316	445	0.71	-0.94	0.93	0.88
30.LAp4g12_EddieandTimmy_Part_D11	327	445	0.73	-1.01	0.87	0.80
31.LAp5g12_EddieandTimmy_Part_D12	368	445	0.83	-1.35	1.04	1.26
32.MAp3g12_AttheFunFair_Part_A1	47	63	0.75	-0.47	1.21	1.11
33.MAp4g12_AttheFunFair_Part_A2	48	63	0.76	-0.53	1.02	1.20
34.MAp5g12_AttheFunFair_Part_A3	42	63	0.67	-0.18	1.01	1.00
35.SCp3g12_LadybugLifeCycle_Part_B4	54	62	0.87	-1.00	0.83	0.52
36.SCp4g12_LadybugLifeCycle_Part_B5	58	62	0.94	-1.42	0.88	0.47
37.SCp5g12_LadybugLifeCycle_Part_B6	40	62	0.65	-0.09	1.19	1.20
38.SSp3g12_MusicInstoftheWorld_Part_C7	59	61	0.97	-1.76	0.97	0.74
39.SSp4g12_MusicInstoftheWorld_Part_C8	43	61	0.70	-0.31	1.00	1.08
40.SSp5g12_MusicInstoftheWorld_Part_C9	41	61	0.67	-0.19	0.83	0.79

Table 4.1.1.1F
Distribution of Mean-Square Fit Statistics: Grades 1–2 Reading

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 "distorting or degrading measurement"	N = 0 % = 0%	N = 1 % = 2.5%
1.5–2.0 "unproductive but not degrading"	N = 0 % = 0%	N = 4 % = 10%
0.5–1.5 "productive for measurement"	N = 40 % = 100%	N = 27 % = 67.5%
< 0.5 "less productive but not degrading"	N = 0 % = 0%	N = 8 % = 20%
Total	N = 40 % = 100%	N = 40 % = 100%

4.1.1.2. Rasch Analyses for the 3–5 Grade-level Cluster

Results of the Rasch item analysis for grades 3–5 Speaking are reported in Table 4.1.1.2A. The p-value and measure columns show that within the two parts of the assessment, each Speaking task was more difficult than the one before it, and the first folder had slightly more students answer tasks correctly than did the more difficult second folder. Such patterns are consistent with the test design.

Table 4.1.1.2B summarizes the infit and outfit findings. All infit statistics are within the productive range and thus suggest that all of the Speaking items are producing good measurement of students' speaking proficiency. For the outfit, 2 items fall within the range considered by Linacre (2002) to be productive for measurement, 3 items are less productive but not degrading, and 3 items are considered to have potentially distorting or degrading measurement. Further examinations of the three items revealed that some students had unusual response patterns, for example, getting all of the items in the first folder correct but all of the items in the second folder wrong. These unlikely response patterns negatively affected the outfit statistics for these items.

Table 4.1.1.2A
Rasch Item Analysis: Grades 3–5 Speaking

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.3562_SIp1g35TinaYangLunch_Part_A_T1	680	684	0.99	-6.36	0.74	0.02
2.3563_SIp2g35TinaYangLunch_Part_A_T2	652	684	0.95	-2.00	1.19	1.43
3.3564_SIp3g35TinaYangLunch_Part_A_T3	609	684	0.89	0.05	1.28	7.40
4.5533_LSp1g35ClassLeaderWind_Part_B_T1	673	684	0.98	-4.22	1.32	0.18
5.5534_LSp2g35ClassLeaderWind_Part_B_T2	646	684	0.94	-1.62	0.73	9.90
6.5535_LSp3g35ClassLeaderWind_Part_B_T3	611	684	0.89	-0.03	0.72	0.29
7.5536_LSp4g35ClassLeaderWind_Part_B_T4	468	684	0.68	4.28	0.58	0.56
8.5537_LSp5g35ClassLeaderWind_Part_B_T5	338	684	0.49	9.91	1.06	9.90

Table 4.1.1.2B
Distribution of Mean-Square Fit Statistics: Grades 3–5 Speaking

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0 "distorting or degrading measurement"	N = 0 % = 0%	N = 3 % = 37.5%
> 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 = 2 % = 25%
< 0.5 "less productive but not degrading"	N = 0 % = 0%	N = 3 % = 37.5%
Total	N = 8 % = 100%	N = 8 % = 100%

Tables 4.1.1.2C presents the Rasch results for the 34 Listening items for the 3–5 grade-level cluster, and Table 4.1.1.2D summarizes the results. The first four items are from Step 1, and the latter items are for the Low, Mid, and High placement levels of Step 2. No items were misfitting in terms of infit. In terms of outfit, all items were productive for measurement except one item with very low outfit that overfit the model.

Table 4.1.1.2C

Rasch Item Analysis: Grades 3–5 Listening

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.A_LAp2g35_Mystery_T1	530	682	0.78	-0.42	1.10	1.11
2.A_LAp3g35_Mystery_T2	648	682	0.95	-2.49	1.03	0.95
3.A_LAp4g35_Mystery_T3	617	682	0.90	-1.67	0.93	0.66
4.A_LAp5g35_Mystery_T4	550	682	0.81	-0.65	0.96	0.93
5.A_2179_SIp1g35_FollowingInstructions_Part_A_T1	85	89	0.96	-2.62	0.79	0.40
6.A_2180_SIp2g35_FollowingInstructions_Part_A_T2	61	89	0.69	-0.53	0.94	0.89
7.A_2182_SIp3g35_FollowingInstructions_Part_A_T3	71	89	0.80	-1.03	0.86	0.78
8.A_LAp1g35_MissingCard_Part_B_T4	78	89	0.88	-1.56	1.12	0.92
9.A_LAp2g35_MissingCard_Part_B_T5	51	89	0.57	-0.13	1.08	1.02
10.A_LAp3g35_MissingCard_Part_B_T6	30	89	0.34	0.59	1.02	1.05
11.A_MAp1g35_SchoolLunch _Part_B_T7	43	89	0.48	0.15	1.10	1.11
12.A_MAp2g35_SchoolLunch _Part_B_T8	85	89	0.96	-2.62	0.91	0.73
13.A_MAp3g35_SchoolLunch _Part_B_T9	23	89	0.26	0.84	1.04	1.19
14.B_4983_MAp1g35_SchoolSupplyStore_Part_A_T1	163	323	0.50	0.89	1.03	1.06
15.B_4984_MAp1g35_SchoolSupplyStore_Part_A_T2	126	323	0.39	1.18	0.92	0.92
16.B_4985_MAp1g35_SchoolSupplyStore_Part_A_T3	94	323	0.29	1.44	1.14	1.15
17.B_SCp2g35_StatesofMatter_Part_B_T4	127	294	0.43	1.12	0.95	0.95
18.B_SCp3g35_StatesofMatter_Part_B_T5	146	289	0.51	0.95	1.00	0.99
19.B_SCp4g35_StatesofMatter_Part_B_T6	176	288	0.61	0.67	0.87	0.80
20.B_SSp2g35_OregonTrail_Part_C_T7	158	281	0.56	0.82	1.04	1.05
21.B_SSp3g35_OregonTrail_Part_C_T8	96	281	0.34	1.37	1.07	1.07
22.B_SSp4g35_OregonTrail_Part_C_T9	172	280	0.61	0.68	1.14	1.23
23.B_LAp3g35_CaseofMissingGlobe_Part_D_T10	537	595	0.90	-0.28	1.02	1.47
24.B_LAp4g35_CaseofMissingGlobe_Part_D_T11	326	595	0.55	1.11	0.99	1.00
25.B_LAp5g35_CaseofMissingGlobe_Part_D_T12	435	595	0.73	0.59	0.92	0.91
26.C_MAp3g35_Trees_Part_A_T1	314	318	0.99	-1.50	1.01	0.98
27.C_MAp4g35_Trees_Part_A_T2	183	318	0.58	1.27	0.93	0.95
28.C_MAp5g35_Trees_Part_A_T3	251	318	0.79	0.65	1.01	0.91
29.C_SCp3g35_Adaptations_Part_B_T4	258	317	0.81	0.55	1.01	0.90
30.C_SCp4g35_Adaptations_Part_B_T5	236	315	0.75	0.80	0.96	0.89
31.C_SCp5g35_Adaptations_Part_B_T6	193	315	0.61	1.18	0.95	0.94
32.C_SSp3g35_Alaska_Part_C_T7	178	315	0.57	1.30	0.95	0.91
33.C_SSp4g35_Alaska_Part_C_T8	92	315	0.29	1.90	1.12	1.26
34.C_SSp5g35_Alaska_Part_C_T9	155	315	0.49	1.46	0.98	0.99

Table 4.1.1.2D

Distribution of Mean-Square Fit Statistics: Grades 3–5 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 = 34 % = 100%	N = 33 % = 97.1%
< 0.5 “less productive but not degrading”	N = 0 % = 0%	N = 1 % = 2.9%
Total	N = 34 % = 100%	N = 34 % = 100%

Table 4.1.1.2E presents the results of the Rasch analyses of the 34 Reading items for the 3–5 grade-level cluster, and Table 4.1.1.2F provides the summary of fit statistics. The first four items are from Step 1, and the latter items are for the Low, Mid, and High placement levels of Step 2. According to the infit, all items fit the Rasch model well and are productive for measurement. For the outfit statistic, 32 items are productive for measurement. Two easy items have a very low outfit statistic, indicating that they provide redundant measurement information.

Table 4.1.1.2E
Rasch Item Analysis: Grades 3–5 Reading

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.A_LAp2g35_CanoeAdventure_T1	654	674	0.97	-1.83	0.94	0.48
2.A_2908_LAp3g35_CanoeAdventure_T2	453	673	0.67	0.20	1.07	1.08
3.A_2910_LAp4g35_CanoeAdventure_T3	448	673	0.67	0.19	1.02	1.00
4.A_2913_LAp5g35_CanoeAdventure_T4	524	673	0.78	-0.19	0.94	0.79
5.A_SIp1g35_NewBook_Part_A_T1	101	109	0.93	-2.11	1.01	0.60
6.A_SIp2g35_NewBook_Part_A_T2	78	109	0.72	-0.93	0.85	0.75
7.A_SIp3g35_NewBook_Part_A_T3	69	109	0.63	-0.66	0.75	0.67
8.A_MAp1g35_PaintedWalls_Part_B_T4	58	108	0.54	-0.38	0.86	0.80
9.A_MAp2g35_PaintedWalls_Part_B_T5	36	109	0.33	0.21	0.93	0.89
10.A_MAp3g35_PaintedWalls_Part_B_T6	28	109	0.26	0.39	1.13	1.33
11.A_3061_SCp1g35_Rocks_Part_B_T7	75	109	0.69	-0.83	1.10	1.11
12.A_3062_SCp2g35_Rocks_Part_B_T8	71	109	0.65	-0.72	0.95	0.84
13.A_3063_SCp3g35_Rocks_Part_B_T9	37	109	0.34	0.23	1.16	1.27
14.B_MAp2g35_Pencils_Part_A_T1	449	472	0.95	-1.12	1.01	1.07
15.B_MAp3g35_Pencils_Part_A_T2	197	472	0.42	0.91	1.03	1.09
16.B_MAp4g35_Pencils_Part_A_T3	336	471	0.71	0.18	0.89	0.84
17.B_3119_SCp3g35_ObservingBirds_Part_B_T4	457	470	0.97	-1.50	0.92	0.55
18.B_3120_SCp3g35_ObservingBirds_Part_B_T5	329	469	0.70	0.15	0.92	0.83
19.B_3122_SCp3g35_ObservingBirds_Part_B_T6	200	469	0.43	0.90	1.11	1.18
20.B_SSp2g35_TheKingdomofMali_Part_C_T7	416	469	0.89	-0.52	0.98	0.85
21.B_SSp3g35_TheKingdomofMali_Part_C_T8	264	469	0.56	0.59	1.02	1.02
22.B_SSp4g35_TheKingdomofMali_Part_C_T9	174	466	0.37	1.02	0.94	0.97
23.B_LAp3g35_Nurses_Part_D_T10	284	561	0.51	0.79	1.15	1.17
24.B_LAp4g35_Nurses_Part_D_T11	292	561	0.52	0.76	1.07	1.12
25.B_LAp5g35_Nurses_Part_D_T12	188	560	0.34	1.18	0.94	0.94
26.B_MAp3g35_LearningaboutWeather_Part_A_T1	95	96	0.99	-1.45	0.97	0.34
27.B_MAp4g35_LearningaboutWeather_A_T2	49	96	0.51	1.13	1.03	0.98
28.B_MAp5g35_LearningaboutWeather_A_T3	62	96	0.65	0.84	0.94	0.91
29.B_SCp3g35_CellsandTheirFunctions_Part_B_T4	50	93	0.54	1.09	0.92	0.87
30.B_SCp4g35_CellsandTheirFunctions_Part_B_T5	81	93	0.87	0.15	0.92	0.88
31.B_SCp5g35_CellsandTheirFunctions_Part_B_T6	19	93	0.20	1.79	0.97	0.89
32.B_SSp3g35_AncientCivilizations_Part_C_T7	64	93	0.69	0.75	0.92	1.06
33.B_SSp4g35_AncientCivilizations_Part_C_T8	38	93	0.41	1.34	1.17	1.31
34.B_SSp5g35_AncientCivilizations_Part_C_T9	49	93	0.53	1.11	0.99	0.95

Table 4.1.1.2F
Distribution of Mean-Square Fit Statistics: Grades 3–5 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 = 34 % = 100%	N = 32 % = 94.1%
< 0.5 "less productive but not degrading"	N = 0 % = 0%	N = 2 % = 5.9%
Total	N = 34 % = 100%	N = 34 % = 100%

4.1.2. Descriptive Statistics for Speaking, Listening, and Reading

For the Listening and Reading sections on MODEL™, raw scores can range from 0–19, depending on the placement levels (Low, Mid, or High) of the student. For Speaking, raw scores can range from 0–8. For quality assurance, researchers at CAL recomputed the test administrators' total raw scores for each student. Descriptive statistics for the Speaking, Listening, and Readings sections for grade-level clusters 1–2 and 3–5 are presented in Tables 4.1.2A and 4.1.2B, respectively.

Table 4.1.2A
Descriptive Statistics for Grades 1–2 by Step and Placement Level

Domain	Step and Placement Level	No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking	-	8	575	0	8	6.57	1.78
Listening	Step 1	4	574	0	4	2.74	1.11
	Step 2: Low	9	68	0	9	6.72	1.66
	Step 2: Mid	12	293	2	12	8.46	1.69
	Step 2: High	12	213	4	12	9.22	1.72
	Step 1 and Step 2: Low	13	68	0	13	8.10	2.12
	Step 1 and Step 2: Mid	16	293	4	16	10.95	2.15
	Step 1 and Step 2: High	16	213	7	16	12.74	1.92
Reading	Step 1	4	571	0	4	3.12	1.07
	Step 2: Low	15	120	0	15	7.73	4.94
	Step 2: Mid	12	391	0	12	8.89	2.19
	Step 2: High	12	60	5	12	9.60	1.90
	Step 1 and Step 2: Low	19	121	0	19	9.38	5.29
	Step 1 and Step 2: Mid	16	391	2	16	12.32	2.54
	Step 1 and Step 2: High	16	60	9	16	13.45	1.99

Table 4.1.2B
Descriptive Statistics for Grades 3–5 by Step and Placement Level

Domain	Step and Level	No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking	-	8	684	0	8	6.84	1.51
Listening	Step 1	4	682	0	4	3.44	0.81
	Step 2: Low	9	89	1	9	5.92	1.58
	Step 2: Mid	12	278	0	11	6.41	2.35
	Step 2: High	12	315	3	12	8.23	1.92
	Step 1 and Step 2: Low	13	89	1	13	8.48	2.26
	Step 1 and Step 2: Mid	16	278	3	15	9.77	2.65
	Step 1 and Step 2: High	16	315	7	16	11.98	2.05
Reading	Step 1	4	674	0	4	3.08	0.97
	Step 2: Low	9	109	1	8	5.07	1.88
	Step 2: Mid	12	470	2	12	7.30	2.11
	Step 2: High	12	93	3	12	7.09	2.17
	Step 1 and Step 2: Low	13	109	2	12	7.09	2.29
	Step 1 and Step 2: Mid	16	472	1	16	10.46	2.59
	Step 1 and Step 2: High	16	93	7	16	10.88	2.30

4.2. Results for Writing

Students' writing responses from the field test were analyzed to determine whether they could be accurately scored using the scoring procedure developed and to produce descriptive statistics of the tasks.

4.2.1. Scoring the Writing Responses

4.2.1.1. The Internal CAL Writing Scoring Meeting

The internal CAL writing scoring meeting was held at CAL on October 1–2, 2009. The following CAL employees participated: Dorry Kenyon, Daniel Ginsberg, Sarika Mehta, Abbe Spokane, Abby Davis, and Stephanie Gibson. Carsten Wilmes, the WIDA Assessment Operations Manager, also attended. 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 (AT) team on materials that would eventually become part of the MODEL™ Training Tool Kit.

The WIDA Consortium's Writing Rubric, shown in Figure 4.2.1.1, 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 WIDA-ACCESS Placement Test (W-APT) ™. The rubric reflects and elaborates the Performance Definitions for the levels of English language proficiency and represents the three criteria linguistic complexity, vocabulary usage, and language control for each proficiency level 1–6.

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 4.2.1.1: 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 3–5. 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 that had been previously scored by MetriTech. 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, and language control. Thus, for example, a response that was judged to be a 4 overall, but that was especially strong in vocabulary usage, would receive a score of 4+. 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 100 samples were selected from the field test. Without looking at other people's ratings, each person scored each sample. This scoring process was repeated for grades 1–2.

After the meeting, the raters' scores were manually typed by two data entry specialists into their own separate spreadsheets in MS Excel. Their data entry in the two spreadsheets was then compared by a CAL research assistant, and any discrepancies were manually corrected.

In preparation for the upcoming external writing scoring meeting in which all students' writing responses would be rated, CAL employees organized the scores from the internal CAL writing scoring meeting from lowest to highest. Papers on which at least three scorers agreed and all scores were within one score above or below the score on which there was most agreement were considered the clearest samples at a score point. Twenty of these papers were chosen for use in calibration sets for the external writing scorers.

4.2.1.2. The External Writing Scoring Meeting

For the external writing scoring meeting, CAL staff recruited external raters via advertisements in various local distribution lists and on Craigslist. Selected participants were sent relevant training materials, including sections of the Test Administration Manual and PowerPoint presentations. Participants were to study the materials and to come prepared to the meeting.

The external writing scoring meeting was held from November 4–5, 2009 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 was interested in observing the efficacy of the self-instructional materials and receiving user feedback on the Writing training materials. CAL employees Dorry Kenyon, Stephanie Gibson, Abby Davis, and Abbe Spokane facilitated the meeting.

On the first day of the meeting, two CAL employees and three external raters were assigned to score the writing samples for grades 1–2, and two CAL employees and six external raters were assigned to score the writing samples for grades 3–5. An additional scoring session was held on November 13 to finish scoring some writing samples for grades 3–5.

At the beginning of the external Writing scoring meeting, CAL facilitators had the raters practice the scoring of prescored writing samples in order to gauge how accurate the scorers would be after self-training. After recording everyone's scores, a discussion about the ratings occurred, 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 papers from the field test. These papers had been selected by CAL staff in advance. (For more information about the selection of the calibration set, see Chapter 4.2.1.1.) 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, and language control. The goal was for raters to achieve an 80-percent adjacent agreement with the scores assigned by CAL. Raters who did not meet this goal for the first calibration set were assigned a second calibration set of 10 papers and again had the percentage of adjacent agreement computed. CAL staff felt that raters were calibrated and could accurately score the remainder of the field test papers when they met the 80-percent adjacent agreement on the calibration sets. At this point, the raters were allowed to begin the scoring of remainder of the sets.

After the external writing scoring meeting had ended, the raters' scores, which had been captured on scannable forms filled out by the raters themselves, were scanned by a CAL research assistant. Each scannable form was scanned with Remark software twice in case the scanner malfunctioned. The data were then cleaned in MS 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 4.2.1.2.

Table 4.2.1.2
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

In order to ensure that each student's Writing papers were accurately scored, a rescoring process was implemented. For each rater, researchers determined if the converted score for Part B or A was higher and kept that score. 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 might show their best abilities. Then researchers compared the converted scores among pairs of raters to determine if rescoring was necessary. If two raters' scores were the same or differed by 1 or 2, the scores were considered to have sufficient agreement. However, if two raters' scores differed by 3 or more, a CAL employee rescored the student's paper. If the CAL rater's score differed from one of the two original scores by 3 or more, another CAL employee rescored again. All scores assigned to the student's Writing papers were retained in the dataset in order to conduct various psychometric analyses.

4.2.2. Descriptive Statistics for Writing

As described in the scoring and rescoring processes above (see Chapter 4.2.1.2), at least two scores were assigned to each student's paper, and up to four scores were assigned to a student's paper if rescoring by CAL raters was needed. Furthermore, more than four scores could be assigned to students' responses that were also part of the calibration set because additional ratings were assigned by CAL raters during the selection of calibration papers and by external raters during their rater training. All scores assigned to the students' papers were retained in the descriptive statistical analyses for the Writing tasks. Converted raw scores ranging from 0 to 18 were used in the analyses.

Because multiple ratings were awarded to each student's paper, many-facet Rasch model was used (Facets software Version No. 3.58.0, Linacre, 2010) in order to take into account the multiple ratings assigned to the students' papers. Specifically, the many-facet Rasch model was used to compute a Fair Average or fair score for each student's paper, to examine task difficulty, and to understand sources of variability in the scores (see Chapter 6.3.4.2 for details about the many-facet Rasch analyses).

Based on the parameters estimated by a two-facet Rasch model, a Fair Average can be derived for each student. Fair Average is the estimated raw score that a particular student's writing paper would have obtained from a scorer of average severity. Fair Averages take into account raters' variation in harshness or leniency, so they are a better representation of student performance on the Writing tasks than are simple averages.

4.2.2.1. Descriptive Statistics for the 1–2 Grade-level Cluster Writing Tasks

The frequency distribution, mean, and standard deviation of the rounded Fair Averages for Writing Task 1 for grades 1–2 is shown in Table 4.2.2.1A, and the statistics for Task 2 are shown in Table 4.2.2.1B. There is a fair amount of spread of scores across the entire raw score

distribution, and the majority of the scores are in the middle range of the distribution for both tasks. These results indicate that the difficulty levels of the Writing tasks are appropriate for the students.

Table 4.2.2.1A
Descriptive Statistics for Writing Fair Averages: Grades 1–2 Writing Task 1

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	1	0.3	0.3
1	9	3.1	3.5
2	20	6.9	10.4
3	5	1.7	12.2
4	22	7.6	19.8
5	52	18.1	37.8
6	41	14.2	52.1
7	51	17.7	69.8
8	44	15.3	85.1
9	21	7.3	92.4
10	7	2.4	94.8
11	7	2.4	97.2
12	5	1.7	99.0
13	1	0.3	99.3
14	2	0.7	100.0
Total	288	100.0	
Mean	6.33		
Standard Deviation	2.49		

Table 4.2.2.1B
Descriptive Statistics for Writing Fair Averages: Grades 1–2 Writing Task 2

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	2	0.8	0.8
1	3	1.2	2.0
2	13	5.2	7.2
3	14	5.6	12.7
4	31	12.4	25.1
5	40	15.9	41.0
6	27	10.8	51.8
7	61	24.3	76.1
8	31	12.4	88.4
9	10	4.0	92.4
10	5	2.0	94.4
11	9	3.6	98.0
12	5	2.0	100.0
Total	251	100.0	
Mean	6.10		
Standard	2.34		

Deviation			

4.2.2.2. Descriptive Statistics for the 3–5 Grade-level Cluster Writing Tasks

Table 4.2.2.2A and Table 4.2.2.2B show the frequency distributions, means, and standard deviations of the rounded Fair Averages for each Writing task for grades 3–5. There is a fair amount of spread of scores across the entire raw score distribution, and the majority of the scores are in the middle range of the distribution for both tasks. These results indicate that the difficulty levels of the Writing tasks are appropriate for the students.

Table 4.2.2.2A
Descriptive Statistics for Writing Fair Averages: Grades 3–5 Writing Task 1

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	1	0.2	0.2
1	5	1.1	1.4
2	9	2.1	3.4
3	11	2.5	5.9
4	22	5.0	11.0
5	29	6.6	17.6
6	47	10.8	28.4
7	55	12.6	41.0
8	79	18.1	59.0
9	52	11.9	70.9
10	51	11.7	82.6
11	31	7.1	89.7
12	19	4.3	94.1
13	13	3.0	97.0
14	6	1.4	98.4
15	6	1.4	99.8
16	1	0.2	100.0
Total	437	100.0	
Mean	8.04		
Standard Deviation	2.82		

Table 4.2.2.2B
Descriptive Statistics for Writing Fair Average: Grades 3–5 Writing Task 2

Converted Raw Score	Frequency	Percent	Cumulative Percent
0	2	0.8	0.8
1	0	0.0	0.8
2	0	0.0	0.8
3	1	0.4	1.1
4	8	3.1	4.2
5	19	7.3	11.5
6	28	10.7	22.1
7	52	19.8	42.0
8	49	18.7	60.7
9	23	8.8	69.5
10	36	13.7	83.2
11	13	5.0	88.2
12	6	2.3	90.5
13	4	1.5	92.0
14	2	0.8	92.7
15	7	2.7	95.4
16	8	3.1	98.5
17	4	1.5	100.0
Total	262	100.0	
Mean	8.50		
Standard Deviation	2.92		

5. Linking MODEL™ to WIDA ELP Levels

As discussed in Chapter 1 of this report, in order to make the scores on MODEL™ more understandable to educators, students' performances on MODEL™ are interpreted in terms of WIDA's English Language Proficiency levels. These score interpretations are presented in the form of lookup tables that show for each grade and domain the WIDA ELP level scores that correspond with students' raw scores and scale scores. This chapter of the report explains how a linking study was conducted to link MODEL™ scores to WIDA ELP levels and how the lookup tables were derived. The MODEL™ lookup tables can be found in the appendices of the *WIDA MODEL™ Test Administration Manual* for grades 1–2 and grades 3–5 (MetriTech and CAL, 2010).

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).

Because ACCESS scores are already linked to the WIDA ELP levels, the most straightforward approach would be to equate performances on MODEL™ to those on ACCESS. However, the data that we had from the field test came from students who were administered ACCESS in the spring of the 2008–2009 school year and who had also taken the new MODEL™ test in September/October 2009. Because of this time lag between the two administrations, strict psychometric equating of the two forms was not possible. As a result, we investigated methods of linking scores on MODEL™ to scores on ACCESS so that they could be interpreted in terms of the ELP Standards while recognizing the following constraints:

- Sample size. Our goal was to test 300 students per level of test form, but because of time limitations and difficulties in securing test sites, we had fewer students for some of the forms (see Tables 3.2C and 3.2D in Chapter 3.2). Approaches to linking had to be robust to smaller sample sizes.
- Item adaptation. Retired folders taken from ACCESS and used on MODEL™ were adapted, in some cases changed minimally (e.g., color added to the pictures) and in some cases changed more drastically (e.g., new graphics or text). At the very least, folders were not in the same context (i.e., sequence in a test booklet, surrounded by the same set of folders) as they appeared on ACCESS.
- Individual student growth in proficiency. While we had scores on the same students for ACCESS and MODEL™, any linking methodology used must acknowledge that the students have likely grown in English language proficiency from the time that the students took ACCESS to the time that they took MODEL™.

In the following sections of this chapter, we discuss the methodology used to determine how performances on MODEL™ could be linked to scores on ACCESS and thereby interpreted in terms of the WIDA ELP levels. For Listening and Reading, we first used a psychometric method and a qualitative method to estimate the difficulty measure of items. We then evaluated those methods by applying those results to the field test data and comparing the resultant growth in student ability to the expected growth based on models created by Gary Cook, Associate Scientist at the Wisconsin Center for Education Research. For Writing and Speaking, qualitative interpretations of performances on MODEL™ were used to establish the scale, as described below.

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

5.1.1. Method 1: Common Item Linking

Common item linking was conducted to place the students' MODEL™ scores on the same scale as ACCESS using items that are common between ACCESS and MODEL™ as anchors. In this procedure, the item difficulty estimates based on the operational ACCESS results were used to fix, or anchor, the difficulty estimates of the same items that also appeared in MODEL™ through

a series of Rasch analyses using Winsteps software (Version No. 3.70.0.5, Linacre, 2011). In the Rasch analyses, items whose empirical difficulty appeared to deviate greatly from ACCESS to MODEL™ were identified using the "displacement" value in Winsteps. These item parameters may not be very stable and therefore may not be appropriate to use as anchors. Using an iterative process, anchored items with absolute displacement value greater than 0.30 were "released"; that is, their difficulty values were no longer pre-set to the item difficulty value obtained through the ACCESS operational analysis. Rather, their difficulty values were estimated using the field test data. This process was repeated until no anchor item showed an absolute displacement greater than 0.30. This final run was used to determine ability estimates for examinees with method 1.

5.1.2. Method 2: Qualitative Estimates (Bookmarking)

To ensure that the final scaling of the Listening and Reading sections of MODEL™ was grounded in an understanding of the WIDA ELP Standards, three CAL employees with experience in applying the Standards participated in a bookmarking study designed to link performance on MODEL™ to the Standards. For the bookmarking study, booklets were created with the items arranged in order of difficulty, from the easiest item to the most difficult. The experts examined these items and, with reference to the performance level descriptors of the Standards, in addition to the Ordered Item Booklets used in the ACCESS Standard Setting Study (Kenyon, 2006), determined the point at which in their estimation a minimally competent student at a given proficiency level would fall below a 50 percent chance of answering an item correctly. Their estimate of the cut score for that proficiency level would then fall somewhere between the difficulty of that item and the next easiest item.

After estimating the cut scores for all proficiency levels, the expert panel discussed their results together. They were then given a chance to adjust their judgments in a second round. The results from the second round were averaged to determine the panel's estimate of the cut scores. The logit values of those cut scores were then plotted against the logit values of the corresponding ACCESS cut scores, and a quadratic regression was conducted to determine the relationship between scores on the two tests. The resulting equation was used to convert the logit values of the MODEL™ items to the ACCESS scale; in turn, those values were used to determine ability estimates for examinees with method 2.

5.1.3. Method 3: Common Person Linking

At this point in the linking study, we had two estimates of student ability: one based on a purely psychometric approach and the other based on a qualitative study. To reconcile the two approaches, we compared the resultant student ability estimates with ability levels predicted by the growth model developed by Gary Cook. Cook's growth model divided the student population into percentiles based on the change in their ACCESS score from one year to the next and their initial proficiency level. For example, a first-grade student from one of the states in the WIDA Consortium who is at proficiency level 4.2 in Listening one year and who improves by 19 scale score points the following year would be in the 50th percentile in terms of gain. We used this model to estimate the average percentile gain of students from the ACCESS administration to the MODEL™ administration based on our estimates of their ability using the two approaches to

linking described above. Because Cook's model is based on growth across a year, and because the time difference between administration of ACCESS and MODEL™ was roughly between six to eleven months, including summer vacation, we expected that the average student gain would be close to the 40th percentile.

5.1.4. Choosing a Linking Method

To determine which values to use, we needed to create a principled procedure for choosing among them. Because the Bookmarking and Common Item procedures were direct analyses of the test forms, we decided to apply both linking procedures and then choose between them. The Common Person procedure would serve as a "tie-breaker," that is, as a way to help choose between the two procedures when we could not make a decision based on our other analyses.

To choose a linking procedure, we required a set of criteria against which to evaluate them. We first looked at what the distribution of proficiency level scores in the field test sample would be based on the Bookmarking results and the Common Item results. If students were accurately placed into the different levels (Low, Mid, and High), we would expect most of them to receive proficiency level scores within the target range of that level. For example, we would expect most students taking the Low-level test form to receive proficiency level scores in the 1–2 range, with some in the 3–4 range, and few or none at the 5–6 range. For this step, "few" was operationalized as 0%–15% of the sample, "some" as 15%–50%, and "most" as over 50%. Table 5.1.4A shows our expectations for each of the three levels. For each of the levels, we established three criteria, each covering a range of proficiency levels. The last column in Table 5.1.4A shows our expectations of the percentage of students taking a test at that level who would fall within that proficiency level range. For example, we would expect more than half of the students taking a Low-level test to be at PL 1 or 2, while less than 15% of those students would be at PL 5 or 6.

Table 5.1.4A
Expected Proficiency Level Distributions by Placement Level

Placement Level	Criterion	Proficiency Level Range	Expectations
Low	1	1–2	>50%
	2	3–4	15%–50%
	3	5–6	<15%
Mid	1	1	<15%
	2	2–5	>50%
	3	6	<15%
High	1	1–2	<15%
	2	3–4	15%–50%
	3	5–6	>50%

We then tested these expectations against the proficiency level distribution of the sample based on the two linking procedures. The results are shown in Table 5.1.4B. For each proficiency level, if the data matched our expectations, we marked it with a "1"; if it did not match, we marked it a "0". Thus, for example, the Reading 1–2 Low test form met our expectations for criteria 1 and 2

using results from both the Bookmarking and the Common Item procedures, but it did not meet our expectations on the third criterion; i.e., with both procedures, the percentage of students at PL 5 or 6 was greater than the expected maximum level of 15%. As seen in Table 5.1.4B, most test forms met all three of our expectations under at least one of the linking procedures.

Table 5.1.4B
Proficiency Level Distribution Results by Placement Level and Linking Method

	Bookmarking			Common Item		
	Criterion			Criterion		
Test Form and Level	**1**	**2**	**3**	**1**	**2**	**3**
Reading 1–2 Low	1	1	0	1	1	0
Reading 1–2 Mid	1	1	1	1	1	0
Reading 1–2 High	1	0	1	1	0	1
Reading 3–5 Low	1	1	1	1	1	1
Reading 3–5 Mid	1	1	1	1	1	0
Reading 3–5 High	1	1	1	1	0	1
Listening 1–2 Low	1	1	1	1	1	1
Listening 1–2 Mid	1	1	1	1	1	1
Listening 1–2 High	1	1	1	1	1	1
Listening 3–5 Low	0	1	0	1	1	1
Listening 3–5 Mid	1	1	1	1	1	1
Listening 3–5 High	1	1	1	1	1	0

It was decided that, to the extent possible, we would use the results from the same method for each test form within a cluster and domain combination. For example, we would want to use the same method for all of the 1–2 Reading test forms. In the case of Reading, the choices were straightforward. Based on the results shown in Table 5.1.4B, we decided to use the Bookmarking procedure for Reading for grades 1–2 and 3–5, because in both cases the Bookmarking procedure met at least as many criteria at each level as the Common Item procedure, and overall the Bookmarking procedure met more criteria on each test form.

For Listening, the decision was more complicated. For the 1–2 test form, all levels met all three criteria using both procedures. Therefore, to decide between them, we looked at the results from the Common Person procedure described in Chapter 5.1.3, with the expectation that the average student gain would be close to the 40th percentile. We calculated the average difference between observed gain using the Bookmarking and Common Item procedures and expected gain in scale scores from ACCESS to MODEL™ at the 40th percentile for each placement level (Low, Mid, and High) in each domain in each cluster. Table 5.1.4C shows these average differences for all test forms. In the table, an average difference closer to zero indicates a closer alignment between the average observed gain using the linking procedure and the average expected gain. Thus, for Listening for grades 1–2, the average observed gain using the Bookmarking procedure was 10.7 points higher than the average expected gain, while the average observed gain using the Common Item procedure was 5.2 points higher than the average expected gain. Because the average results from the Common Item procedure were closer to the 40th percentile than the

results from the Bookmarking procedure, we decided to use the results from the Common Item method for Listening for grades 1–2.

Table 5.1.4C
Average Differences Between Observed Gain and Expected Gain in Scale Score Points by Cluster and Domain

	Reading		Listening	
Linking Method	**1–2 Grade-level Cluster**	**3–5 Grade-level Cluster**	**1–2 Grade-level Cluster**	**3–5 Grade-level Cluster**
Bookmarking	0.6	0.5	10.7	0.4
Common Item	7.7	4.2	5.2	-6.7

Referring to Table 5.1.4B, for the 3–5 Listening form, we found that the Low level met all three of our criteria with the Common Item procedure but not the Bookmarking procedure, while the High level met all three of our criteria with the Bookmarking procedure but not the Common Item procedure. The Mid level met all three criteria using both procedures. Therefore, we decided to go with the results from the Common Item procedure for the Low level and the Bookmarking procedure for the High level. For the Mid level, because both linking procedures performed similarly, we examined the difference between observed gain based on those two procedures and the expected gain based on our Common Person procedure. The differences between observed and expected gain for this Mid level form are shown in Table 5.1.4D. We found that the difference for the Bookmarking procedure was smaller than that for the Common Item procedure (-5.8 scale scores points vs. -12.3 scale score points). Therefore, we decided to use the results from the Bookmarking procedure for the 3–5 Mid-level form.

Table 5.1.4D
Average Differences Between Observed Gain and Expected Gain in Scale Score Points: Grades 3–5 Listening Mid

Linking Method	**Average Difference**
Bookmarking	-5.8
Common Item	-12.3

The information in Table 5.1.4C shows additional evidence to support our choice of linking procedures for the Reading 1–2 and 3–5 forms. For these forms, the average scale scores based on the Bookmarking procedure were closer to the expected gain at the 40th percentile than were the average scale scores based on the Common Item procedure. With both sets of selection criteria suggesting consistent results, we were confident in our choices of linking procedures.

Table 5.1.4E summarizes our choice of linking procedure for each test form and placement level.

Table 5.1.4E
Choice of Linking Procedure by Test Form and Placement Level

Test Form	Placement Level	Linking Procedure Used
Reading 1–2	Low	Bookmarking
	Mid	
	High	
Reading 3–5	Low	Bookmarking
	Mid	
	High	
Listening 1–2	Low	Common Item
	Mid	
	High	
Listening 3–5	Low	Common Item
	Mid	Bookmarking
	High	Bookmarking

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

The MODEL™ Writing and Speaking sections had no tasks in common with the ACCESS Writing and Speaking sections, so qualitative methods were used to determine the interpretation of the MODEL™ scores.

5.2.1. Writing

To link scores on the MODEL™ Writing section to scores on the ACCESS Writing section, an expert panel of three CAL staff members with experience in standard setting for ACCESS examined Writing samples that had been produced by MODEL™ field test students at each raw score point. Using the WIDA ELP performance level descriptors, along with the Writing portfolios used in standard setting for ACCESS (Gottlieb, Cranley, & Oliver, 2007), the members of the panel assigned scale scores to the Writing samples. They then discussed their results together, and they were given a chance to adjust their judgments in a second round. The results from the second round were averaged to determine the panel's estimate of the scale score corresponding to each raw score point.

Table 5.2.1A and Table 5.2.1B show the results of the study, along with the *a priori* proficiency level for the higher grade in the cluster. The raw score is the rating given to the Writing sample by the rater. The *a priori* proficiency levels are the levels that correspond to the raw scores in the first column. The scale score is based on the results of the standard setting study described above. The proficiency levels for each grade are based on those scale scores. Thus, for example, a raw score of 2- is interpreted as an *a priori* proficiency level score of 2.2. When the expert panel reviewed a sample with that raw score, they assigned it a scale score of 253, which corresponds to a proficiency level score of 2.1 for grade 2 and 2.5 for grade 1. Note that no portfolios had raw scores higher than 5 for grades 1–2 or higher than 6 for grades 3–5. Similarly, no portfolio had a raw score lower than 1+ for either grade-level cluster.

Table 5.2.1A

Writing Scale Scores and Proficiency Levels: Grades 1–2

Raw Score	*A priori* Proficiency Level	Scale Score	Grade 2 Proficiency Level	Grade 1 Proficiency Level
1-	1.2	n/a	n/a	n/a
1	1.5	n/a	n/a	n/a
1+	1.8	243	1.8	2.1
2-	2.2	253	2.1	2.5
2	2.5	272	2.6	2.9
2+	2.8	275	2.7	3.1
3-	3.2	295	3.3	3.6
3	3.5	303	3.5	3.9
3+	3.8	315	3.9	4.3
4-	4.2	318	3.9	4.4
4	4.5	332	4.4	4.9
4+	4.8	342	4.8	5.2
5-	5.2	357	5.3	5.8
5	5.5	369	5.9	6.0
5+	5.8	n/a	n/a	n/a
6-	6.0	n/a	n/a	n/a
6	6.0	n/a	n/a	n/a
6+	6.0	n/a	n/a	n/a

Table 5.2.1B

Writing Scale Scores and Proficiency Levels: Grades 3–5

Raw Score	*A priori* Proficiency Level	Scale Score	Grade 5 Proficiency Level	Grade 4 Proficiency Level	Grade 3 Proficiency Level
1-	1.2	n/a	n/a	n/a	n/a
1	1.5	n/a	n/a	n/a	n/a
1+	1.8	284	1.9	2.3	2.6
2-	2.2	288	2.0	2.4	2.8
2	2.5	290	2.1	2.4	2.8
2+	2.8	302	2.5	2.8	3.2
3-	3.2	320	3.0	3.4	3.7
3	3.5	328	3.3	3.6	3.9
3+	3.8	337	3.6	3.9	4.2
4-	4.2	356	4.2	4.5	4.9
4	4.5	366	4.5	4.9	5.3
4+	4.8	369	4.6	4.9	5.4
5-	5.2	376	4.9	5.2	5.7
5	5.5	388	5.3	5.7	6.0
5+	5.8	n/a	n/a	n/a	n/a
6-	6.0	393	5.5	5.9	6.0
6	6.0	395	5.6	6.0	6.0
6+	6.0	n/a	n/a	n/a	n/a

For the most part, the proficiency level score based on the results from the expert panel for the higher grade in each cluster was close to the *a priori* proficiency level. The consistency between the standard setting study and the *a priori* scores confirmed *a priori* assumptions. Because of these unobserved scores, and because the results from the standard setting study were very close to the *a priori* score, it was decided to use the *a priori* proficiency level for the highest grade level in each cluster and to adjust the scale scores accordingly.

5.2.2. Speaking

For the MODEL™ Speaking section, the same procedure that was used to determine Speaking proficiency scores for ACCESS (Kenyon, 2006) was used to determine the interpretation of the scores. Because the tasks for Speaking were written to elicit speech samples at specific, progressively higher proficiency levels, the standard-setting panel for ACCESS decided that, in order to be rated at a given proficiency level, an examinee had to respond successfully to all prompts at that level and below. Thus, 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 should respond successfully to at least six tasks before being rated at proficiency level 2. It was also decided that a perfect score should be rated at proficiency level 6. In the case of MODEL™, with two folders designed to elicit speech at each of the five proficiency levels, a score of 4 is required for examinees to be rated at proficiency level 2. Table 5.2.2A and Table 5.2.2B show the 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 5.2.2A
Speaking Scale Scores and Proficiency Levels: Grades 1–2

Raw Score	Scale Score	Proficiency Level	
		Grade 2	Grade 1
0	179	1.0	1.1
1	202	1.3	1.3
2	224	1.5	1.5
3	258	1.8	1.8
4	302	2.5	2.6
5	330	3.4	3.5
6	350	4.2	4.3
7	371	5.2	5.2
8	386	6.0	6.0

Table 5.2.2B
Speaking Scale Scores and Proficiency Levels: Grades 3–5

Raw Score	Scale Score	Proficiency Level		
		Grade 5	Grade 4	Grade 3
0	183	1.0	1.0	1.0
1	209	1.0	1.0	1.0
2	236	1.3	1.3	1.3
3	273	1.6	1.6	1.6
4	315	2.0	2.2	2.4
5	340	3.0	3.2	3.4
6	358	4.0	4.1	4.2
7	376	5.0	5.2	5.3
8	394	6.0	6.0	6.0

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" (AERA, APA, & NCME, 1999, p. 9). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property; rather, it is an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment.

In the past two decades, argument-based approaches (Kane 1992, 2006) to validation have emerged. The Assessment Use Argument (AUA) (Bachman, 2005) is a conceptual framework consisting of a series of inferences that link the test taker's performance to a claim along with warrants and backing to support the claim. Following Bachman (2005), this chapter of the report presents an assessment use argument and validity argument to link students' scores on MODEL™ to an interpretation of their English language proficiency.

Central to an argument-based approach to the validity of MODEL™ is a clear statement of proposed interpretations of the English language proficiency of students' MODEL™ scores. In the following sections, we describe two claims that were investigated, the specific statements that elaborate and support the claims (***warrants***), the data and evidence collected to support the claims (***backing***), and the methodology used to test the validity of the claims.

6.2. Claim 1: Interpretations of Scores

Claim 1: MODEL™ tests were 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.

The justification for making such an inference is referred to as a ***warrant***: MODEL™ measures all aspects of English language proficiency as defined by the WIDA ELP Standards.

Backing to support this warrant comes from the test development procedures and content validity (Backing 1.1) and from psychometric and statistical analyses for construct validity (Backing 1.2) and concurrent validity (Backing 1.3).

Backing 1.1: Test development procedures provide evidence of the content validity of MODEL™. 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)—which are grounded in scientifically based research on best practices in general, English as a Second Language, and bilingual education—guided the development of test blueprints, task specifications, and ELP measures for MODEL™. Every item and task on MODEL™ was developed to target at least one of the five WIDA ELP Standards. Additional evidence of content validity is provided by a series of qualitative evaluations of MODEL™ test content during the MODEL™ test development process by content experts: the content review (Chapter 2.3), the international perspectives panel (Chapter 2.4), the bias and content review (Chapter 2.5), and cognitive labs (Chapter 2.6). Backing 1.1 is explained in more detail in Chapter 6.2.1.

Backing 1.2: Construct validity for the Speaking, Listening, and Reading sections is evaluated with the use of Rasch analyses (Chapter 4.1). Items that fit the Rasch model are likely to be measuring the intended construct of English language proficiency and to contain little construct irrelevance. Construct validity for the Writing section is evaluated with the use of rater reliability analyses (Chapter 6.3.4) to indicate that raters used the scoring procedures and training materials to render reliable scores to students' writing samples. Backing 1.2 is explained in more detail in Chapter 6.2.2.

Backing 1.3: Concurrent validity provides evidence that MODEL™ should be correlated highly with other measures of the same ability. In a broader sense, concurrent validity can be conceptualized as part of construct validation (Kane, 2006). Pearson correlations were computed between students' scale scores on MODEL™ and their scale scores on ACCESS (Chapter 6.2.3). Pearson correlations were also computed among students' domain scale scores on MODEL™ (Chapter 6.2.3). Backing 1.3 is explained in more detail in Chapter 6.2.3.

Additional backing to support the warrant are presented throughout this technical report to determine the extent to which MODEL™ scores can be interpreted as valid and meaningful indicators of students' English proficiency as defined by the WIDA ELP Standards. In addition to the evidence provided in this chapter, additional validity evidence can be found in the following chapters: Chapter 1 (Background), Chapter 2 (Test Development), Chapter 4 (Field Test Results), and Chapter 5 (Linking MODEL™ to WIDA ELP Levels). As this technical report progresses from chapter to chapter, it moves through phases of the test development cycle. Each chapter of the technical report details the procedures and processes applied in the creation of

MODEL™ as well as the results. Each chapter also highlights the meaning and significance of the procedures, processes, and results in terms of validity and the relationship to the assessment use. The analyses presented in Chapters 6 and 7 add to the perspectives provided in chapters 1 to 5.

6.2.1. Backing 1.1 (Content Validity)

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 MODEL™ were addressed by the test maps (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.

Additional evidence of content validity was provided by a series of qualitative evaluations of MODEL™ test content during the MODEL™ test development process by content experts. The content review (Chapter 2.3), the international perspectives panel (Chapter 2.4), the bias and content review (Chapter 2.5), and cognitive labs (Chapter 2.6) were conducted to help ensure that items contained the appropriate content for a grade level and proficiency level, 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 by the experts ultimately ensured that the content of MODEL™ formed a legitimate basis upon which to validly derive conclusions about students' English language proficiency.

6.2.2. Backing 1.2 (Construct Validity)

One major threat to construct validity is the inevitable inclusion of construct-irrelevant variance. These are variances that are related to sub-dimensions of abilities measured by the test items and are irrelevant to the focal construct.

Rasch models are confirmatory and assume a strong theoretical grounding for item development. Thus, measures that fit our measurement model may be considered, psychometrically speaking, as very strong measures. Rasch analysis is also a powerful tool for evaluating construct validity. The items that do not fit the Rasch model are instances of multidimensionality. The items that fit are likely to be measuring the single dimension intended by the construct. Therefore, misfitting items are indications of construct-irrelevant variance. As presented in Chapter 4.1, for the Speaking, Listening, and Reading sections of 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 the MODEL™ scores represent the construct that the tests were designed to measure.

For Writing, rater reliability analyses (Chapter 6.3.4) suggest that the score variability associated with the raters was minimal and that the scoring procedures and training materials are sufficient for the raters to render reliable Writing scores. Therefore, there is clear evidence supporting the

claim that the construct-irrelevant variance related to scoring the Writing responses was minimized.

6.2.3. Backing 1.3 (Concurrent Validity)

Concurrent validity refers to how well a test correlates with a previously validated measure. Because MODEL™ and ACCESS were developed using the same WIDA ELP Standards and there was previous evidence that ACCESS is a valid measure of students' English language proficiency, we expect MODEL™ scores to correlate with ACCESS scores in order to claim that MODEL™ is a valid measure of students' English language proficiency.

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 0.9–1.0 is considered very high, 0.7–0.9 is high, 0.5–0.7 is moderate, 0.3–0.5 is low, and 0.0–0.3 is little (Hinkle, Wiersma, and Jurs, 1979).

Table 6.2.3A shows the Pearson correlations between MODEL™ scale scores and ACCESS scale scores for the four language domains and the Overall composite for grades 1–2. The correlations range from a minimum of 0.551 for Speaking to a maximum of 0.768 for the Overall composite score. These moderate to high correlations provide support to the claim that MODEL™ assesses the construct of English language proficiency.

Table 6.2.3A
Pearson Correlations: MODEL™ Field Test Scale Scores and ACCESS Operational Test Scale Scores for Grades 1–2

	Speaking	Listening	Writing	Reading	Overall
Pearson Correlation	.551**	.596**	.706**	.565**	.768**
N[9]	468	468	468	468	468

**. Correlation is significant at the 0.01 level (2-tailed).

Table 6.2.3B shows the correlations between MODEL™ scale scores and ACCESS scale scores for the four language domains and the Overall composite for grades 3–5. The correlations for the test for grades 3–5 are somewhat lower, ranging from a minimum of 0.379 for Listening to a maximum of 0.748 for the Overall composite score.

[9] Correlations are calculated only from students who had scores on all four domains in both MODEL™ and ACCESS.

Table 6.2.3B
Pearson Correlations: MODEL™ Field Test Scale Scores and ACCESS Operational Test Scale Scores for Grades 3–5

	Speaking	Listening	Writing	Reading	Overall
Pearson Correlation	.381**	.379**	.661**	.600**	.748**
N	620	620	620	620	620

**. Correlation is significant at the 0.01 level (2-tailed).

A possible explanation for these less-than-perfect correlations is that the two tests were administered at different times—ACCESS in Spring 2009 and MODEL™ in Fall 2009. We cannot expect all students to grow at equal rates in language proficiency, so we cannot expect the correlations to be as strong as they would be if the tests were administered a short time apart.

The correlations among the language domains in MODEL™ were also computed. Because all domain tests in MODEL™ were administered around the same time on the same students, because all domain tests measure closely related constructs, and because general English language proficiency should underlie proficiency in the individual domains, we expect a moderately high correlation between the scale scores in the domains. In particular, we expect related domains such as Speaking and Listening or Reading and Writing to show a relatively high correlation. Other, unrelated domains might not be correlated as strongly.

Table 6.2.3C shows the correlations between the MODEL™ domain scale scores for grades 1–2. Overall, correlations are low to moderate and range from 0.421 between Speaking and Writing to 0.600 between Writing and Reading. As expected, the strongest correlations are between the related domains of Speaking and Listening as well as Reading and Writing.

Table 6.2.3C
Pearson Correlations: MODEL™ Field Test Domain Scale Scores for Grades 1–2

		Speaking	Listening	Writing	Reading
Speaking	Pearson Correlation	1	.574**	.421**	.460**
	N	502	502	502	502
Listening	Pearson Correlation		1	.443**	.539**
	N		502	502	502
Writing	Pearson Correlation			1	.600**
	N			502	502
Reading	Pearson Correlation				1
	N				502

**. Correlation is significant at the 0.01 level (2-tailed).

Table 6.2.3D shows the correlations between the MODEL™ domain scale scores for grades 3–5. Overall, correlations are low to moderate and range from 0.308 between Speaking and Reading to 0.561 between Listening and Reading. As expected, Reading and Writing also have moderately strong correlations.

Table 6.2.3D
Pearson Correlations: MODEL™ Field Test Domain Scale Scores for Grades 3–5

		Speaking	Listening	Writing	Reading
Speaking	Pearson Correlation	1	.491**	.356**	.308**
	N	640	640	640	640
Listening	Pearson Correlation		1	.509**	.561**
	N		640	640	640
Writing	Pearson Correlation			1	.551**
	N			640	640
Reading	Pearson Correlation				1
	N				640

**. Correlation is significant at the 0.01 level (2-tailed).

6.3. Claim 2: Consistency of Scores

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

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

WIDA has strived to reduce the chance of measurement error in the items and test forms by designing tests that contain a large-enough sample of high-quality items in order to better sample students' performance. This ensures that students would receive similar scores on the test over repeated test administrations. However, score consistency is a matter of degree and needs to be examined using empirical data. The degree of score consistency for MODEL™ was examined using measures of test reliability. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) defines reliability as "the consistency of [educational] measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). Analysis of test reliability provides information about the likelihood that students would receive the same score on the test over repeated test administrations.

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

Warrant: MODEL™ tests produce scores that are consistent across different test administrations, different test items and tasks, and (for the performance tasks) different raters.

Backing to support this warrant comes from analyses of MODEL™ test administration and Step 2 placement procedures (Backing 2.1 and 2.2) and from psychometric and statistical analyses of test reliability and rater reliability (Backing 2.3 and 2.4).

Backing 2.1: Test administration procedures are standardized to reduce the chance of measurement error due to differences in administration procedures. Backing 2.1 is explained in more detail in Chapter 6.3.1.

Backing 2.2: Step 2 placement procedures ensure that students are placed in the proper test level. Placing students in the proper test level allows more accurate measurement of students' abilities and lessens measurement error due to floor or ceiling effect. Backing 2.2 is explained in more detail in Chapter 6.3.2.

Backing 2.3: Analysis of test reliability provides support that students would receive similar scores on the test over repeated test administrations (assuming that no additional learning has taken place). Backing 2.3 is explained in more detail in Chapter 6.3.3.

Backing 2.4: Analysis of rater reliability provides support that the rating process and the training materials are working as intended and that the agreements among raters are adequate. Backing 2.4 is explained in more detail in Chapter 6.3.4.

6.3.1. Backing 2.1: Standardized Test Administration Procedures

WIDA has attempted to address environmental factors by specifying to test administrators the room setup, appropriate amounts of light and noise, desk arrangements, duration of testing times, and security of materials, among other things. To minimize differences in administration procedures and in raters' variation on the Writing and Speaking sections, WIDA has produced the following training materials for test administrators: the *WIDA MODEL™ Test Administration Manual* that details how to prepare for, administer, score, and interpret scores on MODEL™; the *WIDA MODEL™ Test Administration Training Video* that covers general information on the structure of MODEL™, includes commentary from the test developers, and shows scenes demonstrating the administration; and the *WIDA MODEL™ Training Toolkit CD-ROM* that introduces the various presentations and resources—PDFs, Excel workbooks, and PowerPoint presentations—available for people preparing to administer MODEL™.

6.3.2. Backing 2.2: Reliability of Step 2 Placement

Placing students in the proper test level allows a more accurate measurement of students' abilities and lessens measurement error due to floor or ceiling effect. 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, which is scored and, together with the Speaking score, determines the placement of Low, Mid, or High for Listening Step 2. In general, the Listening placement algorithms were found to do a good job of directing the field test examinees into a

level of items that was neither too easy nor too difficult. For grades 1–2, no examinees in levels Mid or High got all items incorrect, and very few students in levels Low or Mid got all of the items correct. For grades 3–5, very few examinees in the level Low topped out, and examinees in levels Mid and High had fairly normal distributions of scores with few examinees at either extreme.

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, which is scored and, together with the Writing Quick Score, determines the placement of Low, Mid, or High for Reading Step 2. In general, the Reading placement algorithms were found to do a good job of directing the field test examinees into a level of items that was neither too easy nor too difficult. For grades 1–2, no examinees in levels Mid or High got all items incorrect, and very few students in levels Low or Mid got all of the items correct. For grades 3–5, no examinees in the level Low topped out, and examinees in levels Mid and High had fairly normal distributions of scores with few examinees at either extreme.

Overall, these results suggest that the placement algorithms in the MODEL™ Listening and Reading sections worked well to place students into sets of items that showed good measurement for their Listening and Reading proficiency levels.

6.3.3. Backing 2.3: Reliability of the Overall Composite

Because decisions about students' English language proficiency are made based on the Overall composite score, the reliability of that score is important. For each grade-level cluster, a stratified Cronbach's alpha coefficient (Cronbach, Schönemann, and McKie, 1965) was computed, weighted by the contribution of each domain score 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 first requires the estimate of the reliability of 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 coefficient from GENOVA (Brennan & Crick, 2003) was used.

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.

Cronbach's alpha (Cronbach, 1951) is an estimate of reliability of the internal consistency of test items. It expresses how well the items appear to measure the same construct. Conceptually, it may be thought of as the correlation obtained between performances on two halves of the test, if every possibility of dividing the test items in two were attempted. Thus, Cronbach's alpha may be low if some items are measuring something other than what the majority of the items are measuring. As with any reliability index, it is affected by the number of test items (or test score points that may be awarded). That is, *all things being equal*, the greater the numbers of items of like quality, the higher the reliability.

Cronbach's alpha is also affected by the distribution of ability within the group of students tested. All things being equal, the greater the heterogeneity of abilities within the group of students tested (i.e., the more widely the scores are distributed), the higher the reliability. In this sense, Cronbach's alpha is *sample dependent.* Reliability can be as much a function of the test as of the sample of students tested. That is, the exact same test can produce widely disparate estimates of reliability based on ability distribution of the group of students tested.

The values of Cronbach's alpha can range from 0.00 to 1.00. A Cronbach's alpha of 0.70 is widely considered a cut point at which reliability is adequate (DeVellis, 1991), but a higher Cronbach's alpha of 0.80 is preferred for this type of assessment (Nunnally, 1978).

6.3.3.1. Reliability of the 1–2 Grade-level Cluster Test

For grade-level cluster 1–2, the stratified alphas for the Overall score—that is, the composite of Speaking, Listening, Reading, and Writing—given all possible combinations of Listening and Reading placement levels (Low, Mid, and High) are shown in Table 6.3.3.1A. For Listening and Reading, the reliability estimates were based on both Step 1 and Step 2 items. Speaking and Writing do not have placement levels. Variances of each domain and the variance of the weighted Overall composite were computed based only on students who had scores in all four domains. Table 6.3.3.1A shows that the reliability for the Overall composite reaches the 0.80 criteria expected from this type of test (Nunnally, 1978).

Table 6.3.3.1A
Reliability of Overall Composite for Grades 1–2 by Step 2 Placement Level

Speaking	Listening Placement Level	Reading Placement Level	Writing	Stratified Alpha
-	Low	Low	-	0.80
	Low	Mid		0.81
	Low	High		0.82
	Medium	Low		0.81
	Medium	Mid		0.82
	Medium	High		0.82
	High	Low		0.81
	High	Mid		0.82
	High	High		0.82

The reliabilities for the individual Speaking, Listening, and Reading domains are in Table 6.3.3.1B. Cronbach's alphas for Speaking and one Reading placement level are above 0.70 or 0.80 and thus indicate that the items have good internal consistency (DeVellis, 1991; Nunnally, 1978). A few Listening and Reading placement levels have lower-than-expected Cronbach's alphas, which might be the result of items that have unexpected responses, items that are too easy or difficult, or a lack of heterogeneity in the sample.

Table 6.3.3.1B
Reliability of Domains for Grades 1–2 by Step and Placement Level

Domain	Step and Placement Level	No. of Items	No. of Students	Cronbach's Alpha
Speaking	-	8	575	0.81
Listening	Step 1 and Step 2: Low	13	68	0.56
	Step 1 and Step 2: Mid	16	292	0.36
	Step 1 and Step 2: High	16	213	0.37
Reading	Step 1 and Step 2: Low	19	115	0.91
	Step 1 and Step 2: Mid	16	383	0.63
	Step 1 and Step 2: High	16	60	0.53

The reliability of the field test Writing scores was investigated using generalizability theory (Shavelson and Webb, 1991; Brennan, 2001). Generalizability theory was developed to assess reliability of measurement in presence of multiple sources of error. It provides an analytic procedure to partition total variance in observed scores to two or more sources of variances: in our case, one due to the student and one due to the rater. Generalizability theory also provides a coefficient of generalizability based on a particular measurement design that is analogous to reliability coefficient in Classical Test Theory. The Generalizability (G-) coefficient is defined as the ratio of universe score and observed score variance. The software program GENOVA (Brennan and Crick, 2003) was used to apply the generalizability theory approach to estimate the reliability coefficient.

Data from the internal CAL writing scoring meeting (see Chapter 4.2.1.1) were used in this analysis. Five CAL raters had scored a set of randomly selected student papers for Task 1. These

selected papers are the only ones that were rated by all raters, 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. Then, using the same data from the generalizability study, a decision (D) study was conducted. Because it is expected that each student receive only one rating for a MODEL™ Writing task in the operational testing, a D study was conducted to obtain the reliability coefficient (G-coefficient) based on one single rater.

The results of the D study for grades 1–2 are presented in Table 6.3.3.1C. Task 1 used the IT (Integrated) Standard, which includes Social and Instructional Language (SIL), Language of Language Arts (LoLA), and Language of Social Studies (LoSS). There were 41 calibration papers. The G-coefficient based on one rater (0.87) suggests good reliability associated with the MODEL™ Writing score.

Table 6.3.3.1C
Results of the Decision Study for Writing Grades 1–2

Task Number	Standard	Number of Papers	Number of Raters	G-Coefficient Based on One Rater
1	Integrated (IT)	41	5	0.87

6.3.3.2. Reliability of the 3–5 Grade-level Cluster Test

For grades 3–5, the stratified alphas for the Overall score—that is, the composite of Speaking, Listening, Reading, and Writing—given all possible combinations of Listening and Reading placement levels (Low, Mid, and High) are shown in Table 6.3.3.2A. For Listening and Reading, the reliability estimates were based on both Step 1 and Step 2 items. Speaking and Writing do not have placement levels. Variances of each domain and the variance of the weighted Overall composite were computed based only on students who had scores in all four domains. Table 6.3.3.2A shows good reliability for the Overall composite according to the 0.70 criteria (DeVellis, 1991) and the 0.80 criteria that is expected from this type of test (Nunnally, 1978).

Table 6.3.3.2A
Reliability of Overall Composite for Grades 3–5 by Step 2 Placement Level

Speaking	Listening Placement Level	Reading Placement Level	Writing	Stratified Alpha
-	Low	Low	-	0.88
	Low	Mid		0.88
	Low	High		0.89
	Medium	Low		0.88
	Medium	Mid		0.88
	Medium	High		0.89
	High	Low		0.88
	High	Mid		0.88
	High	High		0.89

The reliabilities for the individual Speaking, Listening, and Reading domains are in Table 6.3.3.2B. Cronbach's alpha for Speaking is above 0.70 or 0.80 and thus indicates that the items have good internal consistency (DeVellis, 1991; Nunnally, 1978). A few Listening and Reading placement levels have lower-than-expected Cronbach's alphas, which might be the result of items that have unexpected responses, items that are too easy or difficult, or a lack of heterogeneity in the sample.

Table 6.3.3.2B
Reliability of Domains for Grades 3–5 by Step and Placement Level

Domain	Step and Placement Level	No. of Items	No. of Students	Cronbach's Alpha
Speaking	-	8	684	0.75
Listening	Step 1 and Step 2: Low	13	89	0.58
	Step 1 and Step 2: Mid	16	278	0.57
	Step 1 and Step 2: High	16	315	0.46
Reading	Step 1 and Step 2: Low	13	108	0.54
	Step 1 and Step 2: Mid	16	463	0.58
	Step 1 and Step 2: High	16	93	0.53

The program GENOVA (Brennan and Crick, 2003) was used to apply the generalizability theory approach to estimate the reliability coefficient.

As with grades 1–2, the reliability analysis for grades 3–5 was conducted using only data from the internal CAL writing scoring meeting. Five raters scored a set of randomly selected student papers for Task 1. The results of the decision (D) study are presented in Table 6.3.3.2C. Task 1 used the IT (Integrated) Standard, which includes Social and Instructional Language (SIL), Language of Language Arts (LoLA), and Language of Social Studies (LoSS). There were 67 calibration papers. The G-coefficient based on one rater (0.77) suggests good reliability associated with the MODEL™ Writing score.

Table 6.3.3.2C
Results of the Decision Study for Writing Grades 3–5

Task Number	Standard	Number of Papers	Number of Raters	G-Coefficient Based on One Rater
1	Integrated (IT)	67	5	0.77

6.3.4. Backing 2.4: Rater Reliability

Establishing rater reliability is an important step toward a reliable and valid assessment of students' writing ability. MODEL™ Writing tasks required students to create a response and raters to judge the quality of the students' responses building on their understanding of the construct and the scoring rubric. This is a very complicated process, and many factors—the ability of the student, the difficulty of the task, the scoring process, the nature of the rating scale, and the way in which a rater applies the rating scale—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, and many-facets Rasch analyses were conducted to examine and understand sources of variability in writing scores.

6.3.4.1. Inter-Rater Reliability

For the Writing scoring, each student's writing paper was initially scored by two different raters. Raters were randomly assigned sets of student papers as the first or second read. Some students' writing paper was scored by more than two different raters if rescoring was required. 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 agreements or disagreements between the same two raters across sets. Rather, they are measures of the degree of agreements or disagreements 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 were the same or differed by one raw score point, 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, 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* and a CAL rater rescored the paper.

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 percentage of good agreement (|D|=0–1), sufficient agreement (|D|=2–3), and discrepant (|D|>3) and the Pearson correlation between scores assigned by the first and second rater are reported for grades 1–2 in Table 6.3.4.1A and for grades 3–5 in Table 6.3.4.1B.

For grades 1–2, the percentage of good or sufficient agreements for Tasks 1 and 2 was very high (96.4% and 99.6%, respectively). Very small percentages (3.6% and 0.4%) of pairs of ratings were discrepant and needed to be rescored by CAL raters. The Pearson correlations between the converted raw score assigned by the first and second rater were 0.76 and 0.84, indicating that the raters were fairly consistent in their scoring.

Table 6.3.4.1A
Inter-Rater Reliability for Task 1 and Task 2 for Grades 1–2

			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	250	18	5.88	2.58	6.16	2.59	72.80%	23.60%	3.60%	0.76**
Task 2	248	18	5.77	2.49	6.03	2.67	83.07%	16.53%	0.40%	0.84**

** Correlation is significant at the 0.01 level (2-tailed).

For grades 3–5, the percentage of good or sufficient agreements for Tasks 1 and 2 was high (88.0% and 83.5%, respectively). Although 12.0% and 16.5%, respectively, of the rating pairs had a discrepancy greater than three raw score points and needed to be rescored by CAL raters, given that the range of a possible score is wide, from 0 to 18, such a finding is nonetheless encouraging. The Pearson correlations between the converted raw score assigned by the first and the second rater were 0.74 for Task 1 and 0.73 for Task 2, indicating that the raters were fairly consistent in their scoring.

Table 6.3.4.1B
Inter-Rater Reliability for Task 1 and Task 2 for Grades 3–5

			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	366	18	8.10	3.22	7.90	3.25	55.46%	32.51%	12.02%	0.74**
Task 2	249	18	7.98	2.66	9.28	3.14	53.81%	29.72%	16.47%	0.73**

** Correlation is significant at the 0.01 level (2-tailed).

Overall, the inter-rater agreement between the first and second raters is good for MODEL™ for both grades 1–2 and 3–5. Furthermore, the Pearson correlations suggest that raters ranked or ordered students in a consistent fashion. These results suggest that the rubric and the training materials are working as intended.

6.3.4.2. Facets Analysis

Many-facets Rasch model (Facets software Version No. 3.58.0, Linacre, 2010) was used to examine the sources of variability for the MODEL™ field test Writing scores. Because each student responded only to one Writing task, it was not feasible to examine score variability associated with the Writing tasks. However, each student's writing response was scored by two or more raters, so it was possible to examine the score variability associated with raters. Many-facets Rasch analysis provides analytical tools to examine whether there are some idiosyncratic rater behaviors, for instance, whether certain raters were more severe or more lenient in scoring certain students' papers. Two research questions were examined in the facets analysis: do raters differ in severity with which they rate examinees, and are there raters who rated examinees inconsistently?

A two-facet Rash model was specified, the *examinee* facet and the *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 examinee 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 examinees' papers to estimate the individual examinee abilities and rater severities.

Scores from the internal CAL writing scoring meeting, rater training and rater qualification, and external writing scoring meeting (see Chapter 4.2.1) were all included in the Facet's analyses in order to fully utilize all available information about students' performance on the Writing tasks.

Table 6.3.4.2.A reports measurements of rater severity and fit statistics for each rater who scored Task 1 for grades 1–2. During the estimation process, unexpected ratings flagged by Facets as not fitting the model were removed as outliers. The total number of ratings used in the final results was 734 based on 288 papers. In the table, the Number of Ratings column shows that each rater rendered a different number of ratings. For example, Rater 1 had 153 ratings while Rater 8 only had 17 ratings. The next column, Severity (logits), shows the rater severity in logits where the higher the logit value, the more severe the rater. The results show that Rater 1 was the most severe and rater 9 was the least severe. The fourth column, Standard Error, shows the standard errors of these severity estimates. The last column, Infit Mean Square, indicates the fit statistics of the nine raters' judgments. The fit statistics provide information for identifying the degree to which each element (in this case, each rater) is observed in the way that is expected by the statistical model. Rater misfit can indicate inconsistent rating behavior or idiosyncratic rating style. Infit mean square values between 0.50 and 1.50 are considered good fit (Smith, 2000.) No raters were identified as misfitting.

Table 6.3.4.2A
Measurements of Rater Severity and Fit: Grades 1–2 Writing Task 1

Rater[10]	Number of Ratings	Severity (logits)	Standard Error	Infit Mean Square
Rater 1	153	2.46	0.13	0.86
Rater 2	60	2.16	0.19	0.68
Rater 3	178	1.13	0.11	0.63
Rater 4	110	1.05	0.15	0.92
Rater 5	33	0.58	0.23	1.25
Rater 6	35	0.37	0.22	0.96
Rater 7	113	0.32	0.14	0.82
Rater 8	17	-0.17	0.35	1.07
Rater 9	35	-0.44	0.23	0.81

Table 6.3.4.2B presents measurements of rater severity and fit statistics for each rater who scored Task 2 for grades 1–2. The number of ratings used in the analysis was 510 ratings based on 251 papers. Raters had different degrees of severity. Rater 1 was the most severe and rater 6 was the least severe. Rater 1 had an infit value less than 0.50, which suggests that this rater's ratings are too predictable. This could be due to a central tendency (i.e., using the middle of the rating scale). Rater 5 had an infit value greater than 1.50, which suggests that this rater, who scored only 7 papers, showed more variation in scoring students' papers than expected.

Table 6.3.4.2B
Measurements of Rater Severity and Fit: Grades 1–2 Writing Task 2

Rater	Number of Ratings	Severity (logits)	Standard Error	Infit Mean Square
Rater 1	123	0.83	0.16	0.47
Rater 2	114	0.48	0.16	0.60
Rater 3	127	-0.52	0.15	0.77
Rater 4	9	-0.58	0.51	0.50
Rater 5	7	-1.00	0.62	1.92
Rater 6	130	-1.31	0.15	0.65

Table 6.3.4.2C presents measurements of rater severity and fit statistics for each rater who scored Task 1 for grades 3–5. The number of ratings used in the analysis was 1,234 based on 437 papers. The results suggest that raters showed different degrees of severity. Rater 1 was the most severe and rater 12 was the least severe. All infit mean square values were between 0.50 and 1.50 except for rater 2. Rater 2 had an infit value less than 0.50, which suggests that this rater showed less variation in scoring students' papers than expected. This could be due to a central tendency or halo effect.

[10] Note that raters' numbers are assigned by order of the severity of ratings rather than by their original IDs.

Table 6.3.4.2C
Measurements of Rater Severity and Fit: Grades 3–5 Writing Task 1

Rater	Number of Ratings	Severity (logits)	Standard Error	Infit Mean Square
Rater 1	10	0.73	0.13	0.92
Rater 2	123	0.46	0.40	0.43
Rater 3	67	-0.29	0.16	1.05
Rater 4	217	-0.50	0.14	0.75
Rater 5	38	-0.53	0.08	0.90
Rater 6	127	-0.59	0.11	0.65
Rater 7	95	-0.74	0.10	0.83
Rater 8	155	-0.83	0.17	0.43
Rater 9	67	-0.87	0.16	1.14
Rater 10	134	-1.01	0.11	0.70
Rater 11	102	-1.45	0.12	0.91
Rater 12	99	-1.84	0.12	0.70

Table 6.3.4.2D presents measurements of rater severity and fit statistics for each rater who scored Task 2 for grades 3–5. The number of ratings used in the analysis was 578 with 262 papers. The raters showed different degrees of severity. Rater 1 was the most severe and rater 10 was the least severe. No raters were found to be misfitting in terms of their performance.

Table 6.3.4.2D
Measurements of Rater Severity and Fit: Grades 3–5 Writing Task 2

Rater	Number of Ratings	Severity (logits)	Standard Error	Infit Mean Square
Rater 1	15	2.50	0.38	1.01
Rater 2	156	0.55	0.12	0.86
Rater 3	35	-0.14	0.20	0.95
Rater 4	22	-1.12	0.30	0.69
Rater 5	32	-1.56	0.27	1.01
Rater 6	156	-1.61	0.12	0.86
Rater 7	19	-1.74	0.33	0.63
Rater 8	25	-3.28	0.29	0.89
Rater 9	96	-5.02	0.13	0.66
Rater 10	22	-5.38	0.31	0.68

These Facets analyses results suggest that raters differed in the severity with which they rated the students' Writing papers, which is expected. However, only one rater was found to have an infit value that was greater than 1.50, indicating inconsistency in scoring students' Writing papers. All other raters were found to have scored the students' Writing papers consistently. Because scores from the internal CAL writing scoring meeting, rater training and rater qualification, and external writing scoring meeting were all included in these analyses, these results suggest that the score variability associated with the raters can be considered minimum and that the scoring procedures and training materials appear to be sufficient for the raters to render reliable Writing scores.

7. Development and Technical Properties of MODEL Screener

As discussed in Chapter 1.4 of this report, MODEL Screener is a shorter version of MODEL™ in that it contains the entire Speaking and Writing sections but many fewer items in the Listening and Reading sections. MODEL Screener uses the same set of materials as MODEL™, as throughout MODEL™ test materials, any folders that are also used for administering MODEL Screener are denoted with an orange bar on the page. MODEL Screener can provide 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 measurement of students' English language proficiency is sacrificed in the interest of the shortened test and shortened test administration time. MODEL Screener cannot provide reliable proficiency level scores for the shortened Listening and Reading domains and, therefore, should not be used to determine amount and type of services or exit criteria. For any of these purposes, scores on MODEL Screener should be considered as only one of several elements in the decision process regarding ELL services.

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. This chapter of the report provides details of how folders were selected from the full MODEL™ for MODEL Screener and then provides Rasch item statistics, along with information on validity and reliability.

7.1. Selection of Folders for the Screener

In developing MODEL Screener for grades 1–2 and 3–5, test developers noticed that the Speaking sections in MODEL™ were already short. They also noticed that shortening the Writing sections in MODEL™ to only Part A in the Screener would not result in a good indicator of academic English language proficiency because Part A was not designed to elicit extended discourse. Therefore, the Speaking and Writing sections of the MODEL Screener were kept identical to those in MODEL™. However, test developers noticed that the Listening and Reading sections in MODEL™ contained many folders—three or four folders each depending on placement level—and could thus be shortened to contain fewer folders in the Screener.

In order to shorten the Listening and Reading sections for each grade-level cluster, test developers automatically included for each domain the Step 1 folder and then set out to choose for each domain one Step 2 folder. Test developers used data from the field test to select the Step 2 folders that would accurately measure students' English language proficiency, especially for higher-proficiency students who would be ready to exit ELL services. The ideal folders would be folders for Mid and High proficiency levels (levels 2–5) so they minimize the chance of a student being exited too early from services and struggling in non-ELL classes.

For each grade-level cluster, Rasch analyses in Winsteps software (Version No. 3.70.0.5, Linacre, 2011) were performed for different scenarios of Mid and High folders. The item

difficulties, frequency distributions of the scores, and fit statistics were scrutinized and compared by a team of researchers to choose the folders that best fit what is required for the Screener.

Table 7.1A 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 1–2. These folders are the same as those denoted by an asterisk in the table in Chapter 2.7.

Table 7.1A
List of Final MODEL Screener Folders for Grades 1–2

Folder Title	Source	WIDA ELP Standard[11]	Folder Tier[12]	Step and Placement Level
Speaking				
Library	Retired 103	SIL	N/A	N/A
Bears Doing Chores Outside	Retired 103	LoLA/LoSS	N/A	N/A
Listening				
Art Class	Retired 103 + one new item	LoLA	B+	Step 1
Making Friends	IWW 2008	LoLA	C	Step 2: Mid and High
Writing				
No Eggs	New, In house	IT	N/A	N/A
Flying Kites	New, In house	IT	N/A	N/A
Reading				
Big Balloons	Retired 103 but changed some	LoLA	B+	Step 1
Musical Instruments of the World	IWW 2008	LoSS	C	Step 2: High

Similar procedures were used to select Screener folders for grades 3–5. Table 7.1B 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 3–5.

[11] As described in Chapter 1.2.1, the acronyms for the WIDA ELP Standards can be written as 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. For Writing, IT indicates an integrated task that includes SIL, LoLA, and LoSS.

[12] The folder tier correlates to the proficiency level of items in a folder rather than to the placement level of the test (Low, Mid, and High). Tier A folders have items at proficiency levels 1, 2, and 3. Tier B folders have items at proficiency levels 2, 3, and 4. Tier C folders have items at proficiency levels 3, 4, and 5. Tier B+ folders include items at proficiency levels 2, 3, 4, and 5. Placement levels Low, Mid, and High were originally named after ACCESS tiers A, B, and C with B+ indicating Step 1.

Table 7.1B
List of Final MODEL Screener Folders for Grades 3–5

Folder Title	Source	WIDA ELP Standard	Folder Tier	Step and Placement Level
Speaking				
Tina Yang Lunch	Retired 103	SIL	N/A	N/A
Ernesto's Classroom	Retired 103	LoLA/LoSS	N/A	N/A
Listening				
Mystery	IWW 2008	LoLA	B+	Step 1
Missing Globe	IWW 2008	LoLA	C	Step 2: Mid and High
Writing				
Family Activities	New, In house	IT	N/A	N/A
Lion and Mouse	Retired 103	IT	N/A	N/A
Reading				
Canoe Adventure	Retired 200	LoLA	B+	Step 1
Nurses	Adapted	LoLA	C	Step 2: Mid and High

7.2. Rasch Analyses for the Listening and Reading Sections of Screener

Table 7.2A presents the results of the Rasch analyses on the 7 Listening Screener items for grades 1–2, and Table 7.2B summarizes the findings. The Screener items were taken by 506 students who were in the Mid and High levels. Note that the measures are the final ones after using the Common Item linking procedure (see Chapter 5 for more information). 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. These fit statistics differ from those presented in Chapter 4.1 because different students took different items depending on their placement level of Low, Mid, or High.

Table 7.2A
Rasch Item Analysis: Grades 1–2 Listening Screener

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.A_222_LAp2g12_ArtClass_T1	370	506	0.73	-0.68	0.97	0.96
2.A_223_LAp3g12_ArtClass_T2	408	506	0.81	-1.22	0.98	0.96
3.A_224_LAp4g12_ArtClass_T3	336	506	0.66	-0.32	1.00	1.00
4.A_LAp5g12_ArtClass_T4	362	506	0.72	-0.68	1.08	1.09
5.B_LAp3g12_MakingFriends_Part_C_T7	366	506	0.72	-0.65	1.01	1.04
6.B_LAp4g12_MakingFriends_Part_C_T8	387	506	0.76	-0.90	0.88	0.80
7.B_LAp5g12_MakingFriends_Part_C_T9	329	506	0.65	-0.25	1.07	1.07

Table 7.2B
Distribution of Mean-Square Fit Statistics: Grades 1–2 Listening Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0	N = 0	N = 0
"distorting or degrading measurement"	% = 0%	% = 0%
> 1.5–2.0	N = 0	N = 0
"unproductive but not degrading"	% = 0%	% = 0%
0.5–1.5	N = 7	N = 7
"productive for measurement"	% = 100%	% = 100%
< 0.5	N = 0	N = 0
"less productive but not degrading"	% = 0%	% = 0%
Total	N = 7	N = 7
	% = 100%	% = 100%

Table 7.2C presents the results of the Rasch analyses on the 7 Reading Screener items for grades 1–2, and Table 7.2D summarizes the findings. The Screener items were taken by 61 students who were in the High level. Note that the measures are the final ones after using the Bookmark linking procedure (see Chapter 5 for more information). Infit and outfit statistics show that five items are productive for measurement and two items are less productive but not degrading. These fit statistics differ from those presented in Chapter 4.1 because different students took different items depending on their placement level of Low, Mid, or High.

Table 7.2C
Rasch Item Analysis: Grades 1–2 Reading Screener

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.2895_LAp3g12_BigBalloons_T1	61	61	1.00	-2.76	0.10	0.06
2.2896_LAp4g12_BigBalloons_T2	58	61	0.95	-1.66	0.56	0.57
3.2897_LAp4g12_BigBalloons_T3	59	61	0.97	-1.25	0.60	0.58
4.2898_LAp5g12_BigBalloons_T4	57	61	0.93	-1.27	0.86	0.87
5.SSp3g12_MusicInstoftheWorld_Part_C7	59	61	0.97	-1.76	0.42	0.31
6.SSp4g12_MusicInstoftheWorld_Part_C8	43	61	0.70	-0.31	1.13	1.14
7.SSp5g12_MusicInstoftheWorld_Part_C9	41	61	0.67	-0.19	1.05	0.67

Table 7.2D
Distribution of Mean-Square Fit Statistics: Grades 1–2 Reading Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0	N = 0	N = 0
"distorting or degrading measurement"	% = 0%	% = 0%
> 1.5–2.0	N = 0	N = 0
"unproductive but not degrading"	% = 0%	% = 0%
0.5–1.5	N = 5	N = 5
"productive for measurement"	% = 71.4%	% = 71.4%
< 0.5	N = 2	N = 2
"less productive but not degrading"	% = 28.6%	% = 28.6%
Total	N = 7	N = 7
	% = 100%	% = 100%

The Rasch results for the Screener Speaking section and the Facets results for the Screener Writing section for grades 1–2 are the same as those presented in Chapter 4, as all Speaking and Writing tasks from the full MODEL™ were used for the Screener.

Table 7.2E presents the results of the Rasch analyses on the 7 Listening Screener items for grades 3–5, and Table 7.2F summarizes the findings. The Screener items were taken by 595 students who were in the Mid or High levels. Note that the measures are the final ones after using the Bookmark linking procedure (see Chapter 5 for more information). Infit and outfit statistics were found to be productive for measurement for all items except for one item that had a less productive but not degrading outfit. These fit statistics differ from those presented in Chapter 4.1 because different students took different items depending on their placement level of Low, Mid, or High.

Table 7.2E

Rasch Item Analysis: Grades 3–5 Listening Screener

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.A_LAp2g35_Mystery_T1	486	595	0.82	0.22	1.03	1.06
2.A_LAp3g35_Mystery_T2	577	595	0.97	-1.19	0.60	0.56
3.A_LAp4g35_Mystery_T3	558	595	0.94	-0.59	0.58	0.48
4.A_LAp5g35_Mystery_T4	504	595	0.85	0.07	0.81	0.76
5.B_LAp3g35_CaseofMissingGlobe_Part_D_T10	537	595	0.90	-0.28	0.80	0.79
6.B_LAp4g35_CaseofMissingGlobe_Part_D_T11	326	595	0.55	1.11	1.07	1.06
7.B_LAp5g35_CaseofMissingGlobe_Part_D_T12	435	595	0.73	0.59	0.93	0.91

Table 7.2F

Distribution of Mean-Square Fit Statistics: Grades 3–5 Listening Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0	N = 0	N = 0
“distorting or degrading measurement”	% = 0%	% = 0%
> 1.5–2.0	N = 0	N = 0
“unproductive but not degrading”	% = 0%	% = 0%
0.5–1.5	N = 7	N = 6
“productive for measurement”	% = 100%	% = 85.7%
< 0.5	N = 0	N = 1
“less productive but not degrading”	% = 0%	% = 14.3%
Total	N = 7	N = 7
	% = 100%	% = 100%

Table 7.2G presents the results of the Rasch analyses on the 7 Reading Screener items for grades 3–5, and Table 7.2H summarizes the findings. The Screener items were taken by 561 students who were in the Mid or High levels. Note that the measures are the final ones after using the Bookmark linking procedure (see Chapter 5 for more information). Six items had productive infit and outfit statistics, and one item had an infit and outfit statistic that was less productive but not degrading. These fit statistics differ from those presented in Chapter 4.1 because different students took different folders depending on their placement level of Low, Mid, or High.

Table 7.2G
Rasch Item Analysis: Grades 3–5 Reading Screener

ITEM NAME	SCORE	COUNT	P-VALUE	MEASURE	IN.MSQ	OUT.MSQ
1.A_LAp2g35_CanoeAdventure_T1	555	561	0.99	-1.83	0.25	0.14
2.A_2908_LAp3g35_CanoeAdventure_T2	398	561	0.71	0.20	0.84	0.81
3.A_2910_LAp4g35_CanoeAdventure_T3	418	561	0.75	0.19	0.84	0.77
4.A_2913_LAp5g35_CanoeAdventure_T4	478	561	0.85	-0.19	0.70	0.64
5.B_LAp3g35_Nurses_Part_D_T10	284	561	0.51	0.79	1.06	1.09
6.B_LAp4g35_Nurses_Part_D_T11	292	561	0.52	0.76	1.01	0.99
7.B_LAp5g35_Nurses_Part_D_T12	188	560	0.34	1.18	0.88	0.85

Table 7.2H
Distribution of Mean-Square Fit Statistics: Grades 3–5 Reading Screener

Range of Mean-Square Fit Statistic	Infit	Outfit
> 2.0	N = 0	N = 0
"distorting or degrading measurement"	% = 0%	% = 0%
> 1.5–2.0	N = 0	N = 0
"unproductive but not degrading"	% = 0%	% = 0%
0.5–1.5	N = 6	N = 6
"productive for measurement"	% = 85.7%	% = 85.7%
< 0.5	N = 1	N = 1
"less productive but not degrading"	% = 14.3%	% = 14.3%
Total	N = 7	N = 7
	% = 100%	% = 100%

The Rasch results for the Screener Speaking section and the Facets results for the Screener Writing section for grades 3–5 are the same as those presented in Chapter 4, as all Speaking and Writing tasks for the full MODEL™ were used for the Screener.

7.3. Descriptive Statistics

Table 7.3A shows the descriptive statistics for the Listening and Reading sections of the Screener for grades 1–2. Also included are the descriptive statistics for the Speaking and Writing sections, which are the same as for Speaking and Writing in the full MODEL™ because the Screener included all Speaking and Writing tasks. Note that the Writing statistics were computed based on students' Fair Averages from Facets (see Chapter 4.2.2 and Chapter 6.3.4.2).

Table 7.3A
Descriptive Statistics for MODEL Screener by Domain and Task for Grades 1–2

Domain		**No. of Items**	**No. of Students**	**Min.**	**Max.**	**Mean**	**Std. Dev.**
Speaking		8	575	0.00	8.00	6.57	1.78
Listening		7	506	0.00	7.00	5.06	1.47
Reading		7	61	3.00	7.00	6.20	0.93
Writing	Task 1	-	288	0.48	14.08	6.33	2.49
	Task 2	-	251	0.06	11.95	6.10	2.34

Table 7.3B shows the descriptive statistics for the Speaking, Listening, Reading, and Writing sections of the Screener for grades 3–5.

Table 7.3B
Descriptive Statistics for MODEL Screener by Domain and Task for Grades 3–5

Domain		No. of Items	No. of Students	Min.	Max.	Mean	Std. Dev.
Speaking		8	684	0.00	8.00	6.84	1.51
Listening		7	595	1.00	7.00	5.75	1.16
Reading		7	561	1.00	7.00	4.66	1.38
Writing	Task 1	-	437	0.68	16.31	8.04	2.82
	Task 2	-	262	0.01	16.86	8.50	2.92

7.4. Validity

As a shortened form of MODEL™, MODEL Screener shares certain items and tasks with the full version of the test and, as a result, has similar evidence of content validity, construct validity, and concurrent validity. Much of the validity evidence for MODEL™ provided in Chapter 6 also applies to the Screener, but additional evidence is presented here.

Although MODEL Screener was conceptualized at a later date than was MODEL™, the Screener items and tasks were nonetheless guided by the test maps that were created for MODEL™. The Screener contains folders that represent all four language domains, are aligned to the WIDA ELP Standards, were field tested, and underwent reviews in the international perspectives panel (Chapter 2.4), bias and content review (Chapter 2.5), and cognitive labs (Chapter 2.6).

As seen in Chapter 7.2, according to the Rasch infit and outfit statistics, no Listening or Reading items on Screener were found to have degrading measurement of the construct of students' English language proficiency.

As seen in Chapter 6.2.3, in general, MODEL™ had moderate to high correlations with the previously validated measure ACCESS. Table 7.4A here presents Pearson correlations of students' scale scores on the Listening and Reading sections of the Screener with their scale scores on the Listening and Reading sections of the full MODEL™ for grades 1–2. The correlations were moderate to high, indicating that items in the Screener measure a similar construct as those in the full MODEL™.

Table 7.4A
Pearson Correlations of MODEL Screener with the Full MODEL™ for Grades 1–2

	Listening	Reading
Pearson Correlation	.802**	.693**
N	506	61

**. Correlation is significant at the 0.01 level (2-tailed).

As seen in Table 7.4B, for grades 3–5, students' scale scores on the Screener folders had high correlations with the students' scale scores on the full MODEL™, indicating that items in the Screener measure a similar construct as those in the full MODEL™.

Table 7.4B
Pearson Correlations of MODEL Screener with the Full MODEL™ for Grades 3–5

	Listening	Reading
Pearson Correlation	.714**	.790**
N	595	561

**. Correlation is significant at the 0.01 level (2-tailed).

7.5. Reliability

As mentioned in Chapter 1, the Overall score on MODEL™ and MODEL Screener is a composite that weights the individual domains differently: 35% Reading + 35% Writing + 15% Listening + 15% Speaking. To obtain the reliability for the Overall scores for grades 1–2 and 3–5, stratified alphas (see Chapter 6.3.3) were computed based on the variances and Cronbach's alphas of the individual domains for students who had completed all four domains.

The reliabilities for the individual domains and the Overall composite for the Screener for grades 1–2 are presented in Table 7.5A. The Cronbach's alpha of 0.81 for Speaking is high, but Cronbach's alphas of 0.41 for the Listening Screener and 0.34 for the Reading Screener do not reach the recommended 0.7 or 0.8 (DeVellis, 1991; Nunnally, 1978). These lower-than-expected reliabilities are a reflection of the low number of items in each section. The G-coefficient of 0.87 for Writing, which is from the Decision Study discussed in Chapter 6.3.3.1, suggests high reliability. Although the Cronbach's alphas for Listening and Reading are low, combined with the other domains, they produce a high Overall composite reliability of 0.79. Such a high Overall reliability is important because the Overall score and corresponding proficiency level determine if a student needs ELL services.

Table 7.5A
Overall Composite Reliability: MODEL Screener for Grades 1–2

Component	No. of Items	No. of Students	Weight	Reliability
Speaking	8	575	0.15	0.81
Listening	7	506	0.15	0.41
Reading	7	61	0.35	0.34
Writing	-	41	0.35	0.87
Overall Composite	-	-	-	0.79

The reliabilities for the individual domains and the stratified alpha for the Overall score for the Screener for grades 3–5 are presented in Table 7.5B. The Cronbach's alpha of 0.75 for Speaking is high, but Cronbach's alphas of 0.39 for the Listening Screener and 0.39 for the Reading Screener do not reach the recommended 0.7 or 0.8 (DeVellis, 1991; Nunnally, 1978). These

lower-than-expected reliabilities are a reflection of the low number of items in each section. The G-coefficient of 0.77 for Writing, which is from the Decision Study discussed in Chapter 6.3.3.2, suggests high reliability. Although the reliabilities for Listening and Reading were low, combined with the other domains, they produce a high Overall composite reliability of 0.75. Such a high Overall reliability is important because the Overall score and proficiency level determine if a student needs ELL services.

Table 7.5B
Overall Composite Reliability: MODEL Screener for Grades 3–5

Component	No. of Items	No. of Students	Weight	Reliability
Speaking	8	684	0.15	0.75
Listening	7	595	0.15	0.39
Reading	7	560	0.35	0.39
Writing	-	67	0.35	0.77
Overall Composite	-	-	-	0.75

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

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

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

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). *ELP standards and resource guide, 2007 edition*. 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.

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. (2010). *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 prekindergarten through grade 5 2007 edition.* Madison, WI: Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2011). 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.