

Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test

Series 602, 2023-2024 Administration
Annual Technical Report No. 20A

Executive Summary and Part 1: Purpose, Design, Implementation

Prepared by Center for Applied Linguistics
Language Assessment Division
Psychometrics and Quantitative Research Team
June 2025

© 2024 Board of Regents of the University of Wisconsin System on behalf of the WIDA Consortium.

The WIDA ACCESS for ELLs Technical Advisory Committee

This report has been reviewed by the WIDA ACCESS for ELLs Technical Advisory Committee (TAC), which comprises the following members:

- Gregory J. Cizek, Ph.D., Guy B. Phillips Distinguished Professor, Educational Measurement and Evaluation, University of North Carolina at Chapel Hill
- Claudia Flowers, Ph.D., Professor, Educational Research, Measurement, and Evaluation, University of North Carolina at Charlotte
- Akihito Kamata, Ph.D., Professor, Department of Education Policy and Leadership, Department of Psychology, Southern Methodist University
- Timothy Kurtz, Teacher (retired), Hanover High School, Hanover, New Hampshire
- Carol Myford, Ph.D., Professor Emerita, Educational Psychology, University of Illinois at Chicago
- Micheline Chalhoub-Deville, Ph.D., Professor, Educational Research Methodology, University of North Carolina at Greensboro

Executive Summary

This is the 20th annual technical report on the ACCESS for ELLs English Language Proficiency Test and the seventh report specific to the online version of the ACCESS for ELLs assessment (ACCESS Online) since the online version was launched.

This technical report is produced as a service to members and potential members of the WIDA Consortium and to support states' submissions for U.S. Department of Education English language proficiency assessment peer review. The technical information herein is intended for use by those who have technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). WIDA also produces an annual *Year in Review Report*, intended for a general audience, for readers who are interested in a nontechnical overview of the 2023–2024 ACCESS assessment.

ACCESS for ELLs is intended to assess reliably and validly the English language development of English learners (ELs) in grades K–12 according to the WIDA 2012 Amplification of the English Language Development Standards Kindergarten–Grade 12 (WIDA, 2012). Results on ACCESS for ELLs are used by WIDA Consortium states for monitoring the progress of students, for making decisions about exiting students from language support services, and for accountability. WIDA additionally provides screening instruments for initial identification purposes; however, decision processes on how these are incorporated into identification decisions are at individual states' discretion.

ACCESS for ELLs assesses students in the four domains of Listening, Reading, Writing, and Speaking, as required by federal law (Elementary and Secondary Education Act of 1965, amended 2015; §1111(b)(1)(F); §1111(b)(2)(G)) and provides composite scores as required by the same statute (§3121).

ACCESS for ELLs Online Series 602 was administered in school year 2023–2024 in 36 states, the Bureau of Indian Education, the Department of Defense Education Activity, the U.S. Virgin Islands, the Northern Mariana Islands, and the District of Columbia for a total of 41 state entities (henceforth “states”).

The final number of students who participated in the Series 602 Online ACCESS tests is 2,179,759. The grade with the largest number of students represented in this report was grade 2 with 247,413 students, while the grade with the fewest number was grade 12, with 88,406 students. Of the participating WIDA states, the state with the largest population of EL students was Illinois, with 244,991 students, while the state with the fewest was Palau, with 525 students.

Based on a comparison with prior years’ numbers of participating students, there is a 9% increase in the student population that participated in ACCESS Series 602 testing than ACCESS Series 601 testing.

ACCESS for ELLs Series 602 was offered in two administrative formats, an online format (grades 1–12) and a paper-and-pencil format (kindergarten–grade 12). The current report (WIDA ACCESS Technical Report 20A) provides technical information pertaining to ACCESS for ELLs Series 602 Online. A second report (WIDA ACCESS Technical Report 20B) provides technical information for the ACCESS for ELLs Series 602 Paper assessment, including the Kindergarten assessment.

Contents

| | |
|---|----|
| 1. Purpose and Design of ACCESS..... | 7 |
| 1.1 Purpose Statement | 7 |
| 1.2 The WIDA Standards | 7 |
| 1.3 The WIDA Proficiency Levels..... | 8 |
| 1.4 Language Domains..... | 10 |
| 1.5 Grade-Level Clusters..... | 10 |
| 1.6 Tiers | 10 |
| 2. Test Development..... | 13 |
| 2.1 Item and Task Design | 13 |
| 2.1.1 Listening Items | 13 |
| 2.1.2 Reading Items | 17 |
| 2.1.3 Writing Tasks..... | 20 |
| 2.1.4 Speaking Tasks | 26 |
| 2.2 Test Design..... | 28 |
| 2.2.1 Listening | 28 |
| 2.2.2 Reading | 32 |
| 2.2.3 Writing..... | 34 |
| 2.2.4 Speaking | 37 |
| 2.3 Test Construction..... | 40 |
| 2.3.1 Item and Task Development..... | 40 |
| 2.3.2 Field Testing..... | 45 |
| 2.3.3 Item/Task Review and Selection..... | 52 |
| 3. Test Administration | 56 |
| 3.1 Test Delivery..... | 56 |
| 3.1.1 Listening and Reading | 56 |
| 3.1.2 Writing..... | 56 |
| 3.1.3 Speaking | 56 |
| 3.2 Operational Administration | 57 |

| | | |
|-------|--|----|
| 3.2.1 | Administering the Test Practice..... | 54 |
| 3.2.2 | Listening Test Administration | 54 |
| 3.2.3 | Reading Test Administration | 55 |
| 3.2.4 | Writing Test Administration | 55 |
| 3.2.5 | Speaking Test Administration | 56 |
| 3.2.6 | Test Security | 56 |
| 3.3 | Fairness and Accessibility | 57 |
| 3.3.1 | Support Provided to All ELs | 57 |
| 3.3.2 | Support Provided to ELs with IEPs or 504 Plans..... | 58 |
| 4. | Scoring | 60 |
| 4.1 | Multiple Choice Scoring: Listening and Reading | 60 |
| 4.2 | Scoring Performance-Based Tasks: Writing and Speaking..... | 60 |
| 4.3 | Writing Scoring Scale | 67 |
| 4.4 | Speaking Scoring Scale | 70 |
| 5. | Summary of Score Reports..... | 73 |
| 5.1 | Individual Student Report..... | 73 |
| 5.2 | Other Reports | 76 |

1. Purpose and Design of ACCESS

1.1 Purpose Statement

The purpose of ACCESS for ELLs is to assess the developing English language proficiency of English learners (henceforth ELs) in grades K–12 in the 41 U.S. states, territories, and federal agencies in the WIDA Consortium, first in the English Language Proficiency Standards (Gottlieb, 2004; WIDA, 2007) and then in the amplified 2012 English Language Development (ELD) Standards (WIDA, 2012). The WIDA ELD Standards, which correspond to the academic language used in state academic content standards, describe six levels of developing English language proficiency and form the core of the WIDA Consortium’s approach to instructing and testing ELs. ACCESS may thus be described as a standards-based English language proficiency test designed to measure ELs’ social and academic language proficiency in English. It assesses social and instructional English as well as the academic language associated with language arts, mathematics, science, and social studies, within the school context, across the four language domains (Listening, Reading, Writing, and Speaking).

Other purposes of ACCESS include:

- Identifying the English language proficiency level of students with respect to the WIDA ELD Standards used in all member states of the WIDA Consortium
- Identifying students who have attained English language proficiency
- Assessing annual English language proficiency gains using a standards-based assessment instrument
- Providing districts with information that will help them to evaluate the effectiveness of their language instructional educational programs and determine staffing requirements
- Providing data for meeting federal and state statutory requirements with respect to student assessment
- Providing information that enhances instruction and learning in programs for ELLs.

ACCESS for ELLs is offered in two formats: ACCESS Online (also referred to as *ACCESS Online*), described in this report, and ACCESS Paper, described in a companion report.

1.2 The WIDA Standards

Five foundational WIDA ELD Standards inform the design, structure, and content of ACCESS for ELLs:

- *Standard 1:* ELLs communicate in English for **Social and Instructional** purposes within the school setting.
- *Standard 2:* ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Language Arts**.
- *Standard 3:* ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Mathematics**.

- *Standard 4:* ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Science**.
- *Standard 5:* ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Social Studies**.

For practical purposes, the five Standards are abbreviated as follows in this report:

- Social and Instructional Language: SIL
- Language of Language Arts: LoLA
- Language of Math: LoMa
- Language of Science: LoSc
- Language of Social Studies: LoSS

Every selected response item and every performance-based task on ACCESS for ELLs targets at least one of these five Standards. In Speaking and Writing tasks, the Standards are combined as follows:

- Integrated Social and Instructional Language (SIL), Language of Language Arts (LoLA), and Language of Social Studi(es) (LoSS): IT (Writing only)
- Language of Math (LoMa) and Language of Science (LoSc): MS (Speaking and Writing)
- Language of Language Arts (LoLA) and Language of Social Studies (LoSS): LS (Speaking and Writing)

The overarching goal of ACCESS Online is to measure the academic English language proficiency of students. Proficiency is measured according to a scale, as defined by the WIDA ELD Standards Framework as comprising five levels of proficiency, which are in turn defined in the performance definitions (WIDA, 2012).

The five WIDA ELD Standards should not be thought of in the same sense as content standards (Allen, Carlson, & Zelenak, 1999); rather, they provide the context for assessing a student's language proficiency in a given domain, so the skills that contribute to academic English language proficiency in a domain are the same across the five ELD Standards. In other words, the construct being measured across the five ELD Standards is the same within a domain.

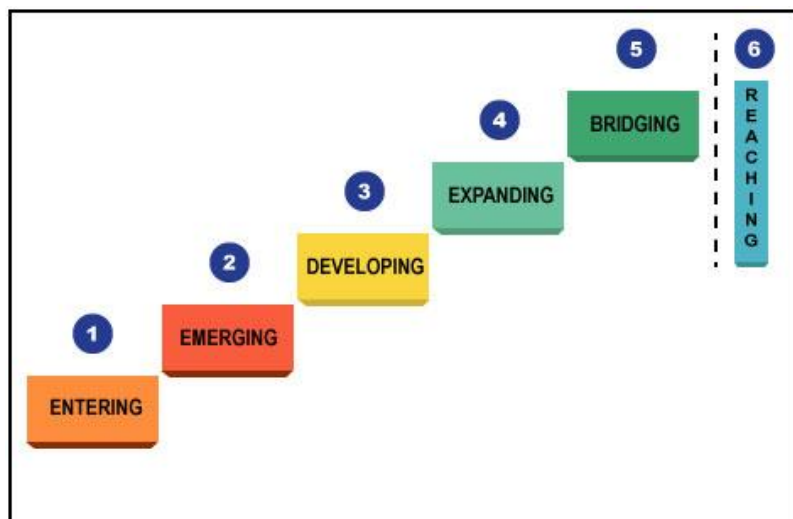
Because of this conceptualization of the WIDA ELD Standards, scores are not reported for each of the Standards, and it is not necessary to assess all five Standards in one domain if each of the Standards is measured on the assessment in some capacity (although ACCESS for ELLs does strive to represent all five WIDA Standards in each domain test).

1.3 *The WIDA Proficiency Levels*

The WIDA ELD Standards describe the continuum of language development via five language proficiency levels (PLs) that are fully delineated in the WIDA ELD Standards document (WIDA, 2012), with scores indicating progression through each level. These levels are *Entering*, *Emerging*, *Developing*, *Expanding*, and *Bridging*. There is also a final stage known as *Reaching*, which is used to describe students who have progressed across the entire WIDA English language proficiency continuum; as this is the end of the continuum, scores do not indicate progression through this level. The proficiency levels are shown graphically in Figure 1.

Figure 1.

The Language Proficiency Levels of the WIDA ELD Standards



These language proficiency levels are embedded in the WIDA ELD Standards in two ways.

First, they appear in the **performance definitions**. The performance definitions describe the stages of language acquisition, providing details about the language that students can comprehend and produce at each proficiency level. The performance definitions are based on three criteria: (1) vocabulary usage at the word/phrase level; (2) language forms and conventions at the sentence level; and (3) linguistic complexity at the discourse level.

Vocabulary usage refers to students' increasing comprehension and production of the technical language required for success in the academic content areas. Language forms and conventions refers to the increasing development of phonological, syntactic, and semantic understanding in receptive skills or control of usage in productive language skills. Linguistic complexity refers to students' understanding or demonstration of oral interaction or writing of increasing quantity and variety.

Second, language proficiency levels are represented through connections to the accompanying **Model Performance Indicators** (MPIs). The MPIs provide a model of the expectations for English learners in each of the five Standards, by grade-level cluster, across the four language domains, for each of the language proficiency levels up to level 5. The grouping of MPIs at proficiency levels 1 through 5 for a given WIDA Standard, grade-level cluster, domain, and topic is called a strand. These MPIs together describe a logical progression and accumulation of skills on the path from the lowest level of English language proficiency to full English language proficiency for academic success. The final level, PL 6: *Reaching*, represents the end of the continuum rather than another level of language proficiency.

Each MPI has a tripartite structure, consisting of a language function, a content stem, and support. The MPIs used on ACCESS can be taken directly from the WIDA English Language Proficiency Standards (WIDA, 2007) or the amplified 2012 ELD Standards (WIDA, 2012). In addition, given that the MPIs in the WIDA Standards are truly "models" and do not cover all possible topics within each Standard for each grade-level cluster and language domain, MPIs

can be “transformed” to accommodate the needs of classroom instruction, as described in the amplified 2012 ELD Standards (WIDA, 2012, p. 11). MPIs are also transformed for the assessment. When MPIs are transformed, one or more of the three aspects of the base MPI are changed. For example, if an MPI from the amplified 2012 ELD Standards (WIDA, 2012) has “categorize” as its language function, it could be transformed to “compare/contrast” or “infer.” Likewise, if the content stem for a grades 9–10 Language of Social Studies strand of MPIs is “supply and demand,” it could be transformed to “freedom and democracy.” Each item specification document for a given WIDA Standard, grade-level cluster, and language domain contains an MPI for each item or task, such that the MPI is the core construct that the given item/task intends to measure. Each selected-response item or performance-based task on ACCESS for ELLs is carefully developed, reviewed, piloted, and field tested to ensure that it allows students to demonstrate accomplishment of the targeted MPI.

In reporting proficiency, WIDA reports scores for each of the domains, in addition to composite scores and an overall score (WIDA, 2021c). So, for each of the domain scores, WIDA reports measures of academic English language proficiency in that domain. More specifically, the score for Speaking is a measure of academic English language proficiency in the domain of Speaking, and likewise for Writing.

1.4 *Language Domains*

The WIDA ELD Standards describe developing English language proficiency for each of the four language domains: Listening, Reading, Writing, and Speaking. Thus, ACCESS for ELLs contains four sections, each assessing an individual language domain.

1.5 *Grade-Level Clusters*

The grade-level cluster structure for ACCESS Online is as follows: 1, 2–3, 4–5, 6–8, and 9–12. Note that the Kindergarten (K) form is not administered online and thus is not covered in this report.

1.6 *Tiers*

ACCESS is designed so that test paths or forms are appropriate to the proficiency level of individual students across the wide range of proficiencies described in the WIDA ELD Standards (Figure 2). Tests must be at the appropriate difficulty level for each individual student to facilitate valid and reliable interpretations of scores. While the grade-level cluster structure is a design feature intended to ensure that the language expectations are developmentally appropriate for students in different age ranges, within each grade-level cluster, students display a range of abilities. Test items and tasks designed for Entering (PL 1) or Emerging (PL 2) students to demonstrate accomplishment of the MPIs at their proficiency level will not allow Expanding (PL 4) or Bridging (PL 5) students to demonstrate the full extent of their language proficiency. Likewise, items and tasks that allow Expanding (PL 4) and Bridging (PL 5) students to demonstrate accomplishment of the MPIs at their level would be far too challenging for

Entering (PL 1) or Emerging (PL 2) students. Items that are far too easy for students may be boring and lead to inattentiveness; items that are far too difficult for students may be frustrating and discourage them from performing their best. But more importantly, items that are too easy or too hard for a student add very little to the accuracy or quality of the measurement of that student's language proficiency.

In the Listening and Reading multistage adaptive tests, students are routed to folders that vary in difficulty, designated as A, B, or C level folders.¹ Tier A folders are intended for students at beginning levels of English language proficiency (PLs 1–3), Tier B folders for students at intermediate levels (PLs 2–4), and Tier C folders for students at more advanced proficiency levels (PLs 3–5). In the domain of Writing, the test forms are designated as either Tier A, which includes tasks written to elicit language up to PL 3, or Tier B/C, which includes tasks written to elicit language up to PL 4 or PL 5. In the domain of Speaking, test forms are designed so that students at very beginning levels of proficiency take a pre-A form, which is designed to elicit language at PL 1; students at early levels of proficiency take the Tier A form, with tasks designed to elicit language at PL 1 and PL 3; and more proficient students take the Tier B/C form, with tasks designed to elicit language at PL 3 and PL 5.

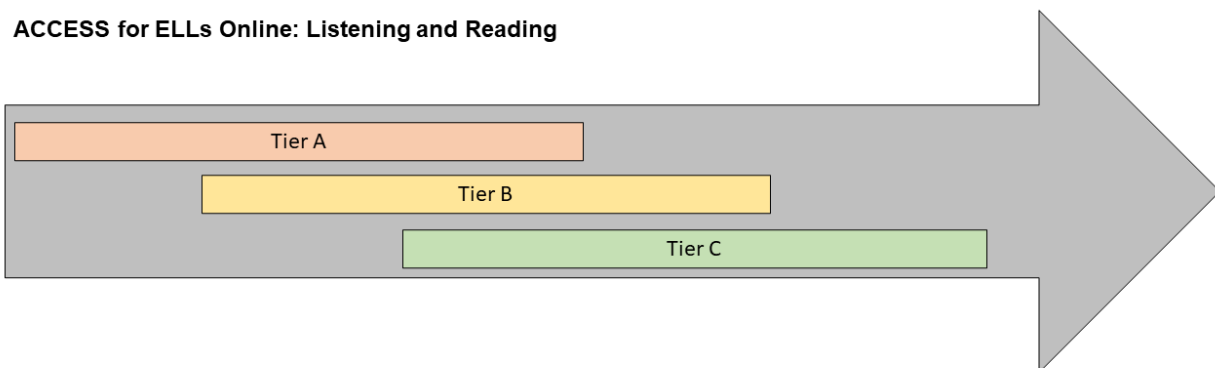
¹ In Listening and Reading, a *Thematic folder*, or folder for short, is a collection of three items constructed around a common theme. For Writing, a thematic folder consists of one or two tasks written to a common theme. For Speaking, a thematic folder consists of two tasks written to a common theme.

Figure 2.

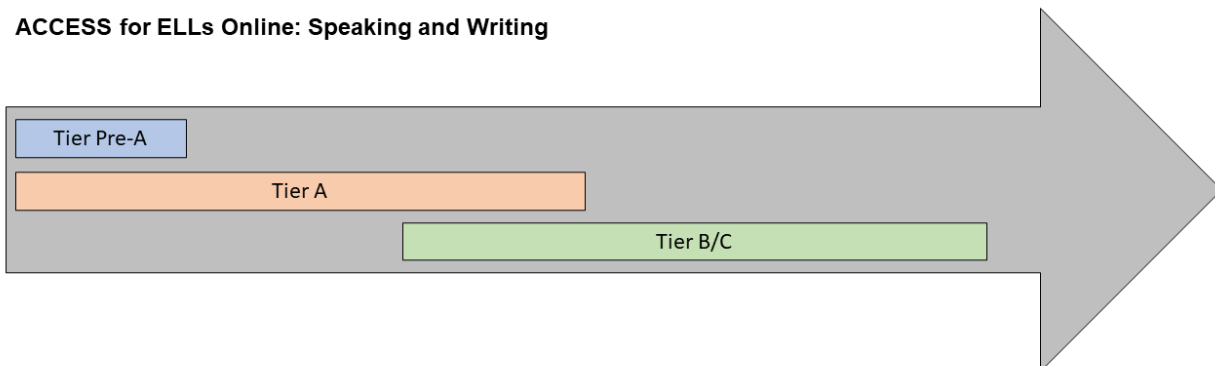
Tiers and Proficiency Levels



ACCESS for ELLs Online: Listening and Reading



ACCESS for ELLs Online: Speaking and Writing



2. Test Development

2.1 *Item and Task Design*

This section describes how the Center for Applied Linguistics (CAL) Test Development (TD) team designs items and tasks to collect the necessary evidence required for the assessment. Items and tasks are discussed by language domain. Readers who are interested in seeing illustrative examples of items and tasks can find these on the [Sample Items page](#) on WIDA's website.

When the task models for ACCESS Online were first developed, CAL and WIDA addressed issues of fairness by ensuring that principles of Universal Design of Assessments (UDA) (National Center on Educational Outcomes, 2021) were adhered to in this design phase. Therefore, CAL, WIDA, and Data Recognition Corporation (DRC) collaborated to design the item and task layout on the screen to be maximally readable/legible with sufficient whitespace, to be accessed intuitively by students, to be accompanied by instructions and practice items to allow students to become accustomed to the test interface, and to contain universal accessibility tools (e.g., magnifier, line guide) as well as tools for accommodation (such as control of test audio and extended response time for the Speaking test). How the CAL TD team ensures fairness by adhering to principles of UDA in item development, in addition to the process by which bias and sensitivity review panels evaluate items and tasks to ensure accessibility and fairness for all students, are described in Section 2.3.1.

2.1.1 Listening Items

All Listening items include a prerecorded stimulus passage and question stem. Listening items are selected-response items, with one key and two distractors as answer choices. Answer choices are primarily graphics (illustrations, photographs, charts/diagrams); for grades 2–12, items that test Listening proficiency at PLs 3–5 may consist of short written text response options that are written to be about two PLs lower than the targeted PL of the Listening item.


Most items on the operational Listening test are traditional multiple choice, though some operational items and some items embedded for field testing purposes may involve enhanced item presentations, including hot spot items (i.e., the student clicks on an area of the screen to respond) and drag-and-drop items (i.e., the student drags an image/text to a specified screen area to respond).


For traditional multiple-choice items, students choose an answer from a set of ordered response options. The response options may be images or text. Students select their answer by clicking anywhere within the box that denotes the response options, including inside the circle that appears to the left of the text or image. Students can change their answer by clicking on a different response option. A screenshot of a sample Listening multiple choice item is provided in Figure 3.

Figure 3.


Multiple Choice Item Layout for the ACCESS Online Listening Test


Listening Practice











1

☐ 

☐ 

☐ 

For hot spot items, students see a large response area. The response area may be an image, a paragraph of text, or some combination of images and text, such as a timeline or a webpage. The answer choices may be pictures or text and are embedded in the response area inside blue boxes. Students answer the question by clicking on one of the boxes in the response area. Each answer choice changes color when selected. Students can change their answers by clicking on a different blue box or by clicking on the reset eraser button, which clears the original response, and clicking on a different blue box. A screenshot of a sample hot spot item is provided in Figure 4.

Figure 4.

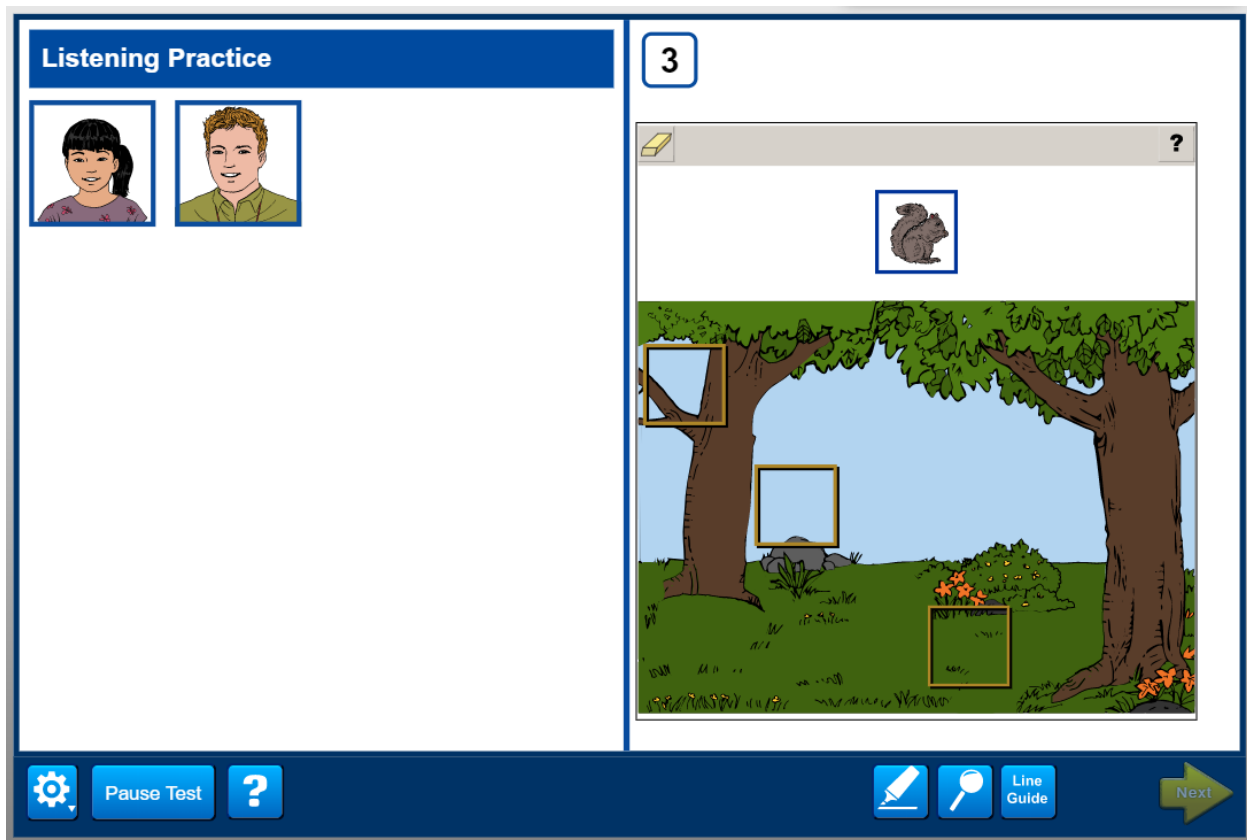
Layout of a Hot Spot Item for the ACCESS Online Listening Test



Drag-and-drop items have two possible formats. In one format, students see one object, either a small image or a line of text, above the response area, which may be an image, a paragraph of text, or some combination of images and text, such as a timeline, a webpage, etc. The response area has three or four blue boxes in it. To show their answer, students click and drag/move the small object into a blue box within the response area. Students do not have to place the object exactly in the blue box; the object snaps into place when students release the mouse button. In this type of drag-and-drop item, students can change their answer by dragging their object into a different blue box in the response area or by clicking on the reset eraser button, which clears the original response, and then dragging the object into a different blue box in the response area. Alternatively, students may see three small objects above the response area. In this case, students select one object to drag into the single blue box within the response area. A screenshot of a sample drag-and-drop item is provided in Figure 5.

Figure 5.

Layout of a Drag-and-Drop Item for the ACCESS Online Listening Test



The number of enhanced items on the Listening test is not specified in the test or item specifications, so the appearance of enhanced items on the test is emergent from the content. In other words, if the content of a given item lends itself well to an enhanced item type, then it is operationalized as such. For example, if an item focuses on three objects contextualized in a larger picture, and the dimensions of the objects do not fit within the standard space for multiple choice response options, like bars in a bar graph, we can operationalize the item as a hot spot item.

Each item on the Listening test targets the language of one of the five WIDA ELD Standards and tests a student's ability to process language at one of the five fully delineated proficiency levels.² Folders group together three test items that are written around a common theme, with each item targeting a progressively higher proficiency level.

- Tier A folders are constructed to target PLs 1 through 3.
- Tier B folders are constructed to target PLs 2 through 4.
- Tier C folders are constructed to target PLs 3 through 5.

² Level 6 is defined as "language that meets all criteria through Level 5, Bridging" and does not have descriptors at the word, sentence, and discourse levels like the other levels.

In the ACCESS Online Listening test, students take a multistage adaptive test form, which routes students to Tier A, B, or C folders as appropriate to their ability level.

Each Listening item appears on its own screen with associated graphic support. Scripts containing the item orientation, stimulus, and question stem are audio recorded with professional voice actors, and a professional recording studio produces the items. Audio playback of test item content is automatic when students advance to the next screen. Listening test content is played one time for students unless the student has a predetermined accommodation allowing for a single repetition of the item stimulus and question stem. Further detail on accommodations can be found in Section 3.3.2.

2.1.2 Reading Items

Reading items are similar in format to Listening items. The stimulus and question stems for Reading items are written text, and answer choices are also primarily written text, though response options for items targeting PLs 1 and 2 may be graphics (illustrations, photographs, charts/diagrams) or text. As with Listening items, Reading items are grouped into thematic folders of three test items each.

- Tier A folders target PLs 1 through 3.
- Tier B folders target PLs 2 through 4.
- Tier C folders target PLs 3 through 5.

In the ACCESS Online Reading tests, students take a multistage adaptive test form, which routes them to Tier A, B, or C folders as appropriate to their ability level.


Most items on the operational Reading test are traditional multiple choice. A screenshot of a sample Reading multiple choice item is provided in Figure 6.

Figure 6.

Multiple Choice Item Layout for the ACCESS Online Reading Test

Reading Practice

Robert, Ava, and Mr. Green are reading about fish.




1




What are they reading about?

☐ Trees

☐ Fish

☐ Birds

As with the Listening test, some operational items and some items embedded for field testing purposes involve enhanced item presentations, including hot spot and drag-and-drop items. The layouts of the Reading hot spot and drag-and-drop items are presented in Figure 7 and Figure 8 respectively.

Figure 7.

Layout of a Hot Spot Item for the ACCESS Online Reading Test

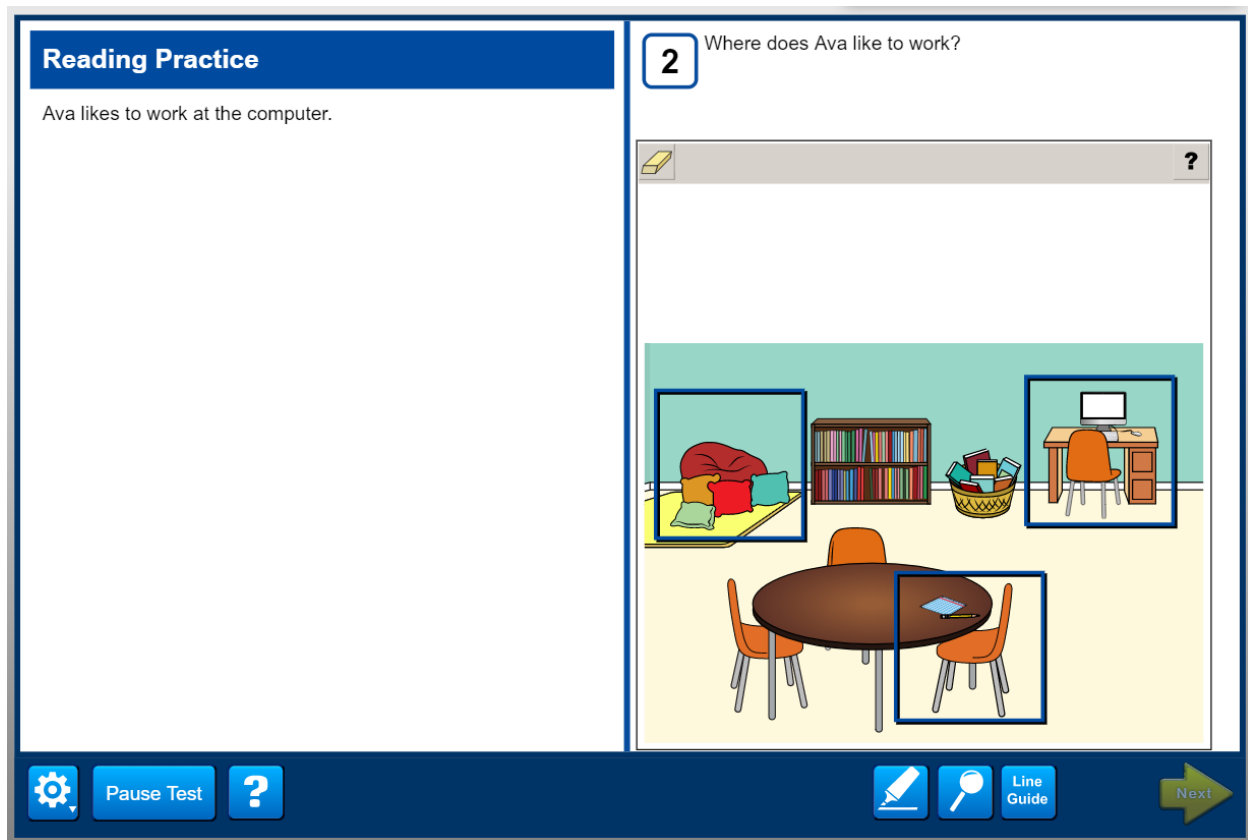


Figure 8.

Layout of a Drag-and-Drop Item for the ACCESS Online Reading Test



The number of enhanced items on the Reading test is not specified in the test or item specifications, so the appearance of enhanced items on the test is emergent from the content. In other words, if the content of a given item lends itself well to an enhanced item type, then it is operationalized as such. For example, if an item focuses on three objects contextualized in a larger picture, and the dimensions of the objects do not fit within the standard space for multiple choice response options, like bars in a bar graph, we can operationalize the item as a hot spot item.

Items have one key and either two or three distractors, depending upon the grade-level cluster and the targeted proficiency level. For grades 1 and 2–3, all items have a key and two distractors. For grades 4–5, 6–8, and 9–12, items targeting PLs 1 and 2 have a key and two distractors, and items targeting PLs 3, 4, and 5 have a key and three distractors. These design decisions were made based on considerations related to reducing cognitive processing load for younger students and lower proficiency level students.

2.1.3 Writing Tasks

Writing tasks are designed to elicit language corresponding to one or more of the WIDA ELD Standards. Tasks appearing on the Tier A test form are designed to allow students to produce writing samples that fulfill linguistic expectations up to PL 3. DRC raters score students' written

responses to these tasks using the entire breadth of the scoring scale(see Section 2.2.3). Therefore, students may achieve proficiency levels higher than PL 3, although the tasks are not designed to elicit extended responses, so the scores are limited by task design. Tasks appearing on the Tier B/C form are designed to give students the opportunity to produce writing samples that fulfill linguistic expectations up to PL 5. Again, although these tasks are designed to elicit extended responses, DRC raters score the responses using all nine categories of the scoring scale, so students' actual performance may extend above or below the PL 5 range.

For students in grades 1–3, the Writing test is not administered via computer. For students in these grades, the test administrator reads from a script and the students respond in a printed test booklet. CAL and WIDA made this design decision when ACCESS Online was first developed, based on the challenge that students at this age have with keyboarding their responses, as CAL and WIDA observed in cognitive labs. Figure 9 provides an example of the paper test booklet, and Figure 10 provides an example of the accompanying script.

Figure 9.

Example Test Booklet for the ACCESS Online Writing Test, Grades 1–3




| | |
|---|--|
| <p style="text-align: right;">Name: _____</p>  <p>What do you see in the picture?</p> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>1 <u>clock</u></p> <p>2 _____</p> </div> <div style="width: 45%;"> <p>3 _____</p> <p>4 _____</p> </div> </div> <div style="text-align: right;">  </div> <p style="font-size: small;">Grade 1, Tier A Sample Item: Indoor Play © 2018 Board of Regents of the University of Wisconsin System</p> | <p style="text-align: right;">Name: _____</p> <p>What is happening in the picture?</p> <p>5 <u>The teacher gives a ball</u> <u>to the girl.</u></p> <p>6 _____</p> <p>7 _____</p> <div style="text-align: right;">  </div> <p style="font-size: small;">© 2018 Board of Regents of the University of Wisconsin System Grade 1, Tier A Sample Item: Indoor Play</p> |
|---|--|

Figure 10.

Example Script for the ACCESS Online Writing Test, Grades 1–3

| | |
|--|---|
| <p>Look at the page with the picture.</p> <p>Scan the room and make sure all students are in the right place.</p> <p>This picture shows students playing indoors.</p> <p>Let's read the question under the picture. It says,</p> <p>"What do you see in the picture?"</p> <p>Look at the picture again. Find the clock and point to it.</p> <p>Scan the room to make sure all students are in the right place.</p> <p>Now find number 1. Number 1 says "clock."</p> <p>Now point to something else in the picture. Write that word next to number 2.</p> <p>Pause while students answer number 2. Monitor students for signs that they understand the task. Answer questions.</p> <p>Next to number 3 and number 4, write the names of two more things that you see in the picture. Spell the best you can. When you finish number 4 and get to the stop sign, put your pencil down and look at me.</p> <p>Do you have any questions?</p> <p>Answer questions.</p> <p>You may begin.</p> <p>Monitor the students. Check to make sure everyone is following directions. If any student is struggling, point to the boy, girl, teacher, ball, goal, net, jump rope, cloud, or rain, and say: <i>What is this?</i> Wait for the student's response, and then say: <i>Now write that on the line.</i></p> <p>Allow a reasonable amount of time for everyone to attempt to write something. Go to the next part when all students have finished writing or about 5 minutes have passed.</p> <p>If some students are still writing, say: <i>Please finish what you are writing now.</i> PAUSE 15 SECONDS.</p> <p>Now look at the top of the next page.</p> | <p>This question at the top of the page says,</p> <p>"What is happening in the picture?"</p> <p>Scan the room and make sure all students are in the right place.</p> <p>In this part, you will write about what the class is doing. Number 5 is done for you. Let's read this sentence together. Put your finger on each word as we read.</p> <p>Make sure students are pointing to the first word in the sentence before reading aloud.</p> <p>"The teacher gives a ball to the girl."</p> <p>Look back at the picture. Find the teacher giving a ball to a girl. PAUSE.</p> <p>Scan the room and make sure all students are in the right place.</p> <p>What else is happening in the picture?</p> <p>Allow time for student response.</p> <p>Now write two sentences about what you see happening in the picture. Write one sentence next to number 6 and one sentence next to number 7. Spell the best you can.</p> <p>Do you have any questions?</p> <p>Answer questions.</p> <p>When you finish and get to the stop sign, put your pencil down and look at me.</p> <p>You may begin.</p> <p>Monitor the students. Check to make sure everyone is following directions. Encourage any struggling students by pointing to one of the groups in the picture and saying: <i>Look at these students. What are they doing?</i> Wait for the student's response. Then point to number 6 and say: <i>Now write that next to number 6.</i></p> <p>Allow a reasonable amount of time for everyone to attempt to write something. End the test when all students have finished writing or about 10 minutes have passed.</p> <p>If some students are still writing, say: <i>Please finish what you are writing now.</i> PAUSE 15 SECONDS.</p> <p>End the testing session by saying:</p> <p>Please put down your pencils, and I will come around to collect your writing.</p> |
| <p>Grade 1, Tier A Sample Item: Indoor Play</p> <p>© 2018 Board of Regents of the University of Wisconsin System</p> | <p>© 2018 Board of Regents of the University of Wisconsin System</p> <p>Grade 1, Tier A Sample Item: Indoor Play</p> |

For students in grades 4–12, writing prompts appear on the computer screen. In the spirit of providing maximal support and making every provision to ensure that students are given the opportunity to demonstrate the full extent of their English language proficiency, modeling is sometimes used to make task expectations as clear as possible to students. For example, the first of a series of questions may already be partially completed, or a sentence starter may be provided. In addition to the task screens, all tasks on the ACCESS Online Writing test contain one or more orientation screens, which introduce the students to the context of the task and provide stimuli to serve as input to the tasks. Figure 11, Figure 12, and Figure 13 show the screen layouts for the tasks on the computer-delivered Writing test for Tier A and Tier B/C, respectively.

Figure 11.

Example Orientation Screen for the ACCESS Online Writing Test, Grades 4–12, Both Tiers

Carlos's Rainy Day

Here is a story about Carlos. He is getting ready for school.

1

2

3

4

7:00 AM

Pause Test ?

Line Guide


Next

Figure 12.


Example Layout for the ACCESS Online Writing Test, Grades 4–12, Tier A

Carlos's Rainy Day


1

A man in a red shirt and blue pants stands in a closet, looking at a yellow jacket hanging on a rack. A blue suitcase is on the floor.


2

The man is now wearing the yellow jacket and is standing in a doorway, looking out. A blue suitcase and a red umbrella are on the floor.

3

The man is walking outside in the rain, holding a red umbrella. He is wearing the yellow jacket and blue pants.

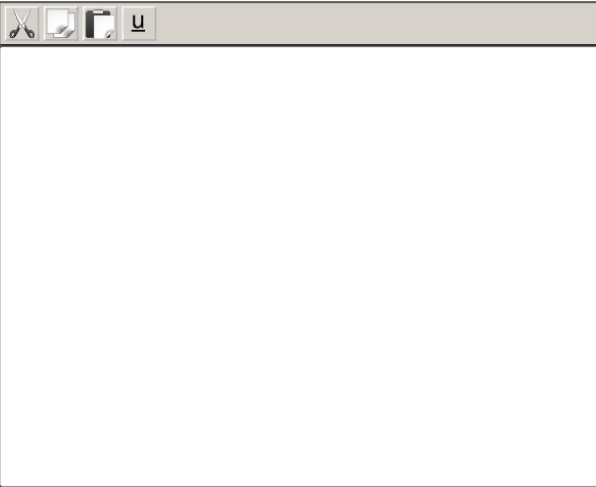
4




The man is standing outside in the rain, talking to a woman who is holding a blue bucket and a hose. They are in front of a brick building.

window

1

What do you see in the pictures? Make a list.

A large white rectangular area for writing. Above it is a toolbar with icons for erasing, copying, pasting, and underlining.











Figure 13.

Example Layout for the ACCESS Online Writing Test, Grades 4–12, Tier B/C

The screenshot displays the ACCESS Online Writing Test interface for the 'Modern Architecture' task. The interface is divided into several sections:

- Task Title:** Modern Architecture
- Navigation Tabs:** Kenzo Tange, Luis Barragan, Plan Your Writing, Check Your Writing.
- Kenzo Tange (1913–2005) Information:**
 - Education:**
 - Architecture degree from Tokyo Imperial University
 - Completed additional studies in city planning
 - Work Experience:**
 - Designed many public buildings in Japan and around the world
 - Led the design of a museum and park used by many people
 - Worked on a plan to help the city of Tokyo grow and change
 - Design Inspiration and Ideas:**
 - Designed buildings to allow for expansion without rebuilding
 - Combined traditional Japanese styles with modern styles
 - Awards and Accomplishments:**
 - Won the Pritzker Architecture Prize in 1987
 - Lectured at many universities
 - Wrote influential books about architecture and urban planning
- Images:**
 - A government office building designed by Tange
 - A museum designed by Tange

The right side of the interface features a large writing area with a toolbar at the top containing icons for cutting, copying, pasting, and underlining. Above the writing area, a task instruction box reads: "1 Write an essay arguing which architect's work was more important. Support your choice using details about both people." The bottom of the interface includes a navigation bar with buttons for settings, pausing the test, help, line guide, back, and next.

Students in grades 4–5 provide either handwritten or keyboarded responses, with the default response mode determined in advance at the state or district level. For students in grades 6–12, keyboarding is the default response mode, with a handwriting option offered as an accommodation.


For students who respond by handwriting in a writing response booklet, the test tasks have a slightly different appearance on the screen when compared to the tasks experienced by students who keyboard their responses. As shown in Figure 14, instead of a writing response space on the right side of the screen, an image of the test booklet appears on the screen to indicate to students where in their writing response booklet they should write.

Figure 14.

Example Layout for an ACCESS Online Tier A Writing Task With the Handwritten (HW) Response Mode

Writing Practice

This picture shows a classroom.



A girl is...


1

What is happening in the picture? Write 1 sentence.


Writing Practice




1

2



Pause Test





Line Guide

Back

Next

2.1.4 Speaking Tasks

Stimuli on the Speaking test include graphics, audio, and text. All stimuli are presented by a Virtual Test Administrator (VTA). The VTA serves as a narrator who guides students through the test and acts as a virtual interlocutor. The VTA is introduced to students during the test directions to establish the testing context.

Task modeling is an essential component of the Speaking test design. In addition to the VTA, students are introduced to a virtual model student during the test directions. Prior to responding to each task, students first listen as the model student responds to a parallel task. The purpose of the model is to demonstrate task expectations to both students and to DRC raters, who score all Speaking task responses.

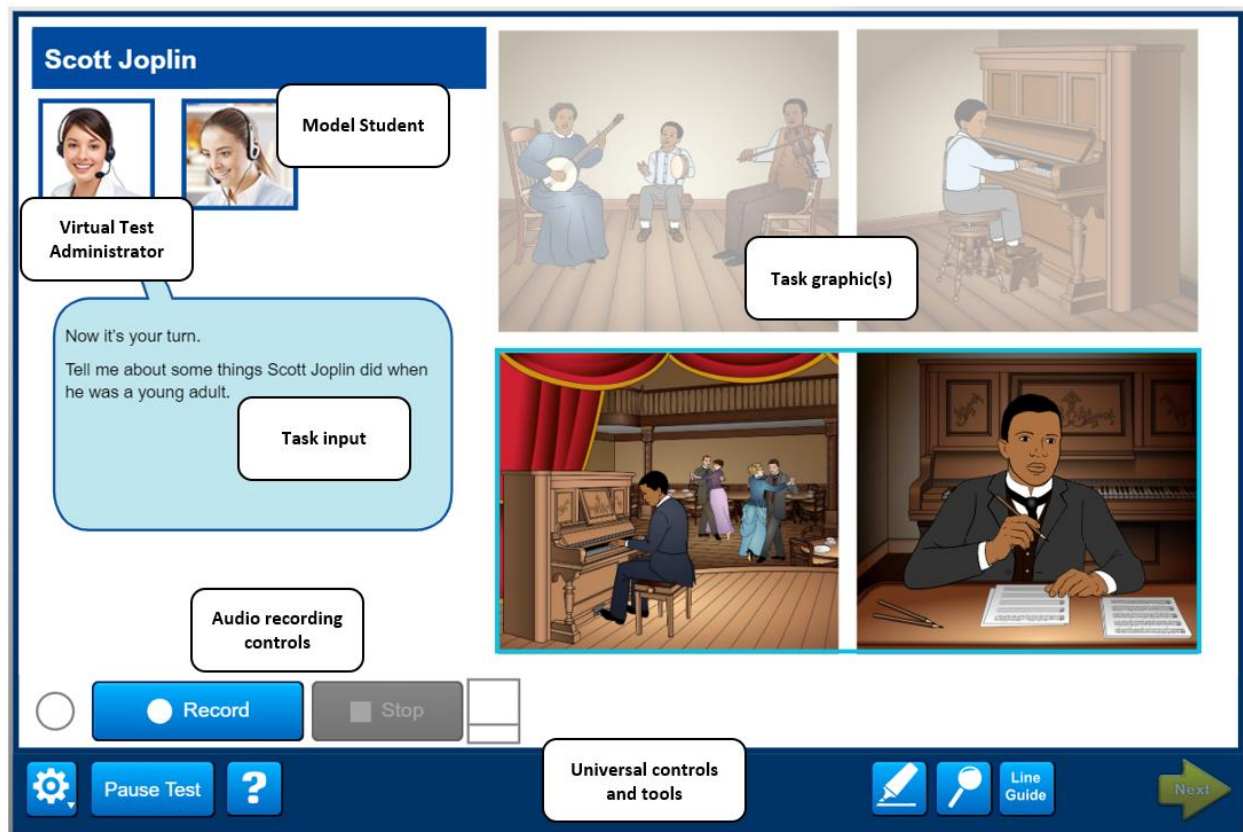
Students navigate through the Speaking test independently and at their own pace. They must listen to all audio on a screen before the test allows them to advance to the next screen. Most students can only listen to the audio stimuli once, although students with a specific accommodation related to audio stimuli may listen to the audio as many times as they wish. The amount of time that students are allowed for recording their responses varies by grade-level

cluster and the target proficiency level of the task; tasks targeting a higher proficiency level are permitted more recording time.³ The amount and complexity of task input varies by grade-level cluster and task level. The purpose of the input is to provide academic content for students to draw on in their responses.

Figure 15 shows the general screen layout of the Speaking test.

Figure 15.

Visualization of the Speaking Test Screen Layout



Both the VTA and the model student are represented within the testing interface by static images. They are portrayed wearing computer headsets with microphones to reflect the actual testing scenario. Test input and stimuli are presented both aurally and in speech bubbles on the screen. Students respond orally to the tasks, with their responses recorded and transmitted to DRC for later scoring.

³ During the piloting of the Speaking test design before ACCESS Online was operational, the response recording time was one of the variables investigated. CAL and WIDA jointly determined the recording times. These times were a compromise between the minimum and maximum times considered. This allows for more time than minimally necessary, while not allowing so much time that students who have already provided a sufficient response feel the need to fill all of the available time.

All Speaking tasks for a given grade-level cluster and WIDA Standard are designed in terms of *panels*; a panel is a thematically related set of three tasks, targeting the elicitation of PL 1, PL 3, and PL 5 language. When the tasks are field-tested, the panels are split out into folders, with each folder containing one or two tasks. Tier Pre-A folders contain a single task targeting PL 1; Tier A folders contain two tasks targeting PL 1 and PL 3; and Tier C folders contain two tasks targeting PLs 3 and 5. For a given pair of Tier A and Tier C folders based on a single panel, the PL 3 task is identical in both folders (see Figure 17 in Section 2.2.4 for an illustration).

2.2 Test Design

This section describes how ACCESS Online is assembled to ensure that the evidence collected is (1) sufficient to make the required decisions based on the test results, and (2) appropriate for the student's level of proficiency. To tailor the test closely to student ability levels while still including items and tasks that assess all the Standards, adaptivity has been built into the test. The Listening and Reading tests both use a multistage adaptive test design. The Writing and Speaking tests are tiered, and placement into the tiers depends on performance on the Listening and Reading tests.

For all four domains, the test design is broken into different tiers (as described in Section 1.6) and stages (as described in this section). For each tier and stage within a given grade-level cluster, a single folder is earmarked for that "slot" on the test. Items selected for each slot must meet strict criteria (in terms of difficulty) to be placed in that slot. This ensures that the item pool is adequate to support the multistage administrations, including the adaptive component in Listening and Reading.

2.2.1 Listening

For the ACCESS Listening test, Table 1 shows, for each grade-level cluster and tier pool, the number of items, the targeted range of WIDA proficiency levels, the proportion of items by item type, the response format, and the scoring procedure.

Table 1.**Number and Types of Items on the ACCESS 602 Listening Test**

| Grade-Level Cluster | Tier Pool | Number of Items | Targeted PL Range | Multiple Choice % | Drag-and-drop % | Hot Spot % | Response Format | Scoring Procedures |
|----------------------------|------------------|------------------------|--------------------------|--------------------------|------------------------|-------------------|-------------------------------|---------------------------|
| 1 | Entry | 6 | PL1-PL4 | 83% | 0% | 17% | Dichotomous selected response | Machine scored |
| 1 | A | 12 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 1 | B | 18 | PL2-PL4 | 77% | 6% | 17% | Dichotomous selected response | Machine scored |
| 1 | C | 18 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | A | 12 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | B | 18 | PL2-PL4 | 78% | 0% | 22% | Dichotomous selected response | Machine scored |
| 2-3 | C | 18 | PL3-PL5 | 89% | 0% | 11% | Dichotomous selected response | Machine scored |
| 4-5 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 4-5 | A | 12 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 4-5 | B | 18 | PL2-PL4 | 83% | 0% | 17% | Dichotomous selected response | Machine scored |
| 4-5 | C | 18 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 6-8 | Entry | 6 | PL1-PL4 | 83% | 0% | 17% | Dichotomous selected response | Machine scored |
| 6-8 | A | 12 | PL1-PL3 | 83% | 0% | 17% | Dichotomous selected response | Machine scored |

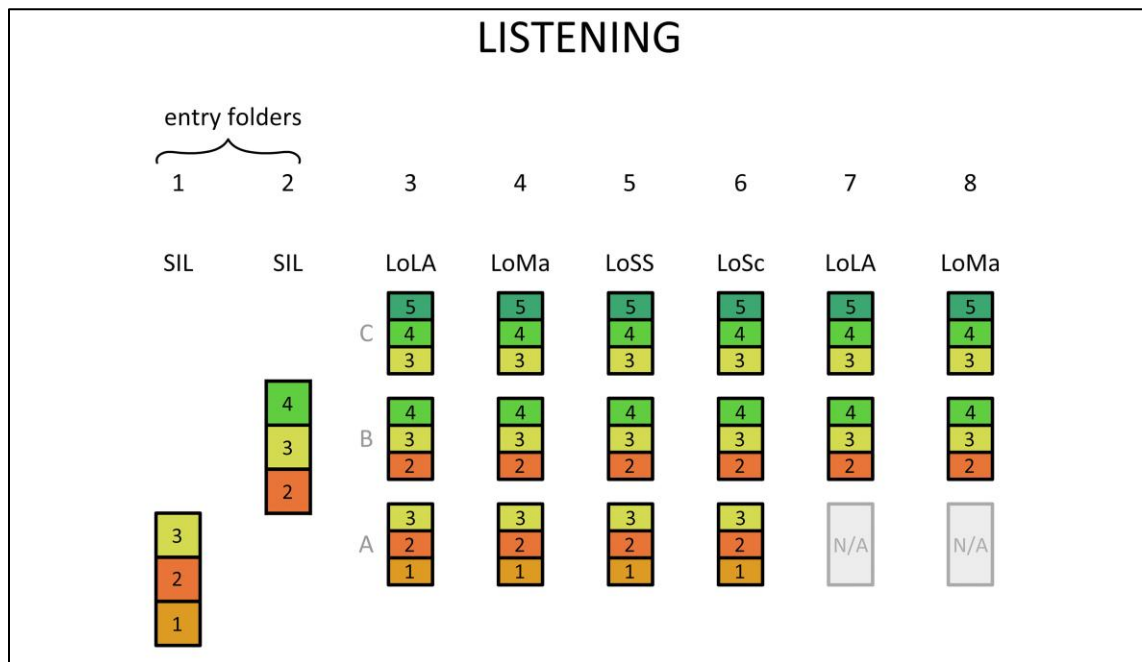
| Grade-Level Cluster | Tier Pool | Number of Items | Targeted PL Range | Multiple Choice % | Drag-and-drop % | Hot Spot % | Response Format | Scoring Procedures |
|---------------------|-----------|-----------------|-------------------|-------------------|-----------------|------------|-------------------------------|--------------------|
| 6–8 | B | 18 | PL2–PL4 | 77% | 6% | 17% | Dichotomous selected response | Machine scored |
| 6–8 | C | 18 | PL3–PL5 | 89% | 0% | 11% | Dichotomous selected response | Machine scored |
| 9–12 | Entry | 6 | PL1–PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 9–12 | A | 12 | PL1–PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 9–12 | B | 18 | PL2–PL4 | 83% | 0% | 17% | Dichotomous selected response | Machine scored |
| 9–12 | C | 18 | PL3–PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |

The Listening test uses a multistage adaptive design, as illustrated in Figure 16. All students begin the Listening test with two entry folders (with three items each) at Stage 1 and Stage 2, both targeting Social and Instructional Language (see Section 1.2 for the WIDA ELD Standards). After a student takes the items in the first two stages of the Listening test, the test engine estimates the student’s ability based on the student’s performance on those six items, and the engine then uses that ability estimate to determine which of the three leveled Language of Language Arts folders in Stage 3 is administered next. Students whose ability estimate predicts a PL score of 5.0 or higher are routed into the folder at the highest level (C in Figure 16); students whose ability estimate predicts a PL score of 2.5 or lower are routed into the folder at the lowest level (A in Figure 16); all others are routed into the B folder.

Throughout the test, the test engine re-estimates a student’s underlying measure of ability with the completion of each folder, and the engine then uses that information to choose the level of the next folder to be administered, following the decision rules above. Thus, each student will trace a tailor-made path through the test according to ability level, but the order of the stages is invariant across students. In total, there are eight possible stages, but a student whose ability estimate falls below PL 2.5 after the sixth stage ends the test at that point. This shortening of the test for students at the lower proficiency levels allows them to demonstrate what they know without subjecting them to additional content, when their ability is not near the cut point where the EL reclassification decision is made. The intent of this design is to ensure coverage of the Standards while delivering a test that closely matches the student’s PL, thus minimizing measurement error. Although timing guidance is included in the test administrator manual (WIDA, 2021a), the Listening test is untimed.

Figure 16.

Format of the Listening Test



2.2.2 Reading

For the ACCESS Reading test, Table 2 shows, for each grade-level cluster and tier pool, the number of items, the targeted range of WIDA proficiency levels, the proportion of items by item type, the response format, and the scoring procedure.

Table 2.

Number and Types of Items on the ACCESS 602 Reading Test

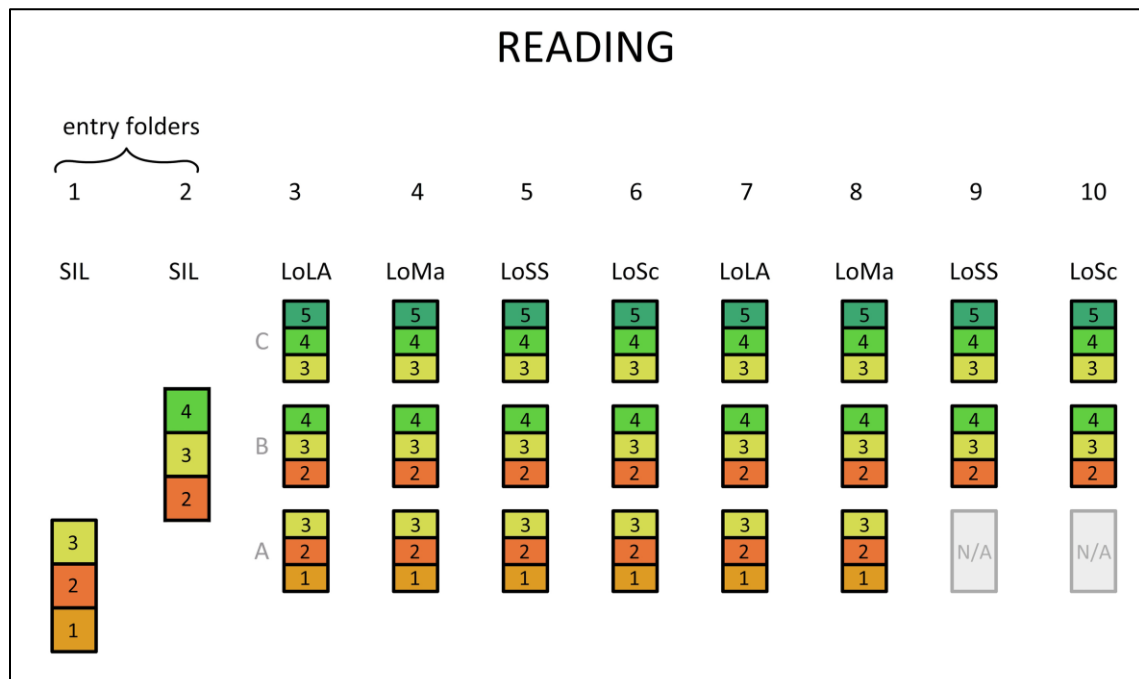
| Grade-Level Cluster | Tier Pool | Number of Items | Targeted PL range | Multiple Choice % | Drag-and-drop % | Hot Spot % | Response Format | Scoring Procedures |
|---------------------|-----------|-----------------|-------------------|-------------------|-----------------|------------|-------------------------------|--------------------|
| 1 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 1 | A | 18 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 1 | B | 24 | PL2-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 1 | C | 24 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | A | 18 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | B | 24 | PL2-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 2-3 | C | 24 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 4-5 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 4-5 | A | 18 | PL1-PL3 | 94% | 0% | 6% | Dichotomous selected response | Machine scored |
| 4-5 | B | 24 | PL2-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |

| Grade-Level Cluster | Tier Pool | Number of Items | Targeted PL range | Multiple Choice % | Drag-and-drop % | Hot Spot % | Response Format | Scoring Procedures |
|----------------------------|------------------|------------------------|--------------------------|--------------------------|------------------------|-------------------|-------------------------------|---------------------------|
| 4-5 | C | 24 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 6-8 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 6-8 | A | 18 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 6-8 | B | 24 | PL2-PL4 | 96% | 4% | 0% | Dichotomous selected response | Machine scored |
| 6-8 | C | 24 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 9-12 | Entry | 6 | PL1-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 9-12 | A | 18 | PL1-PL3 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 9-12 | B | 24 | PL2-PL4 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |
| 9-12 | C | 24 | PL3-PL5 | 100% | 0% | 0% | Dichotomous selected response | Machine scored |

Figure 17 shows the format of the Reading test. The format and adaptivity are like those of the Listening test, but the Reading test consists of 10 stages rather than eight. This reflects the greater weight given to Reading in calculating the composite scores (see Part 2, Chapter 3, Analyses of Composite Scores), as well as the view that literacy skills are paramount in developing academic language proficiency. The greater weight afforded to Reading and Writing resulted from a policy decision by the WIDA Board before the first operational administration of ACCESS. A student whose ability estimate falls below PL 2.5 after the eighth stage ends the test at that point. Although timing guidance is included in the test administrator manual, the Reading test is untimed.

Figure 17.

Format of the Reading Test



2.2.3 Writing

For the ACCESS Writing test, Table 3, shows, for each grade-level cluster and tier, the number of tasks, the targeted range of WIDA proficiency levels, the task type, the response format, and the scoring procedure.

Table 3.

Number and Types of Tasks on the Writing Test

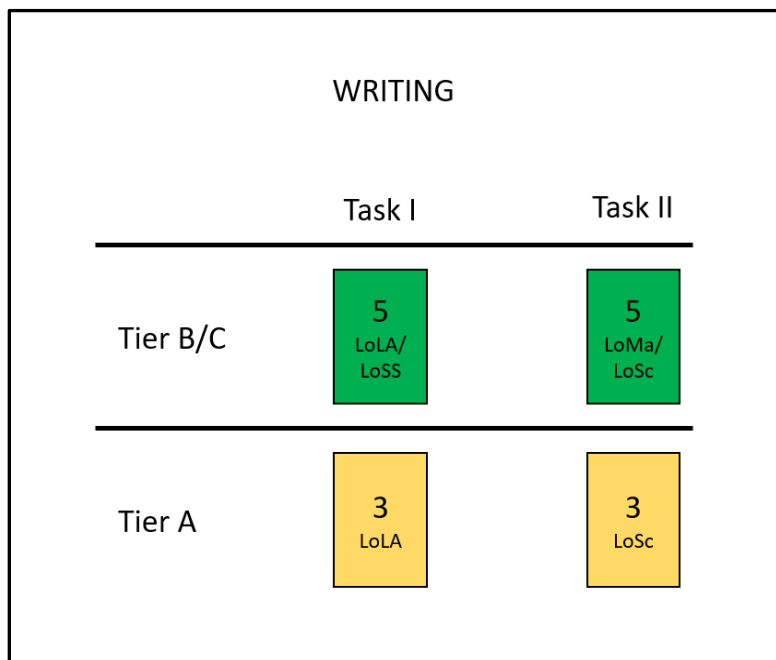
| Grade-Level Cluster | Tier | Number of Tasks | Targeted PL Range | Task Type | Response Format | Scoring Procedures |
|---------------------|------|-----------------|-------------------|------------------------------|--|---------------------------------------|
| 1 | A | 2 | PL1–PL3 | Writing constructed response | Polytomous constructed response; handwritten in test booklet | Human scored: centrally scored by DRC |
| 1 | B/C | 2 | PL2–PL5 | Writing constructed response | Polytomous constructed response; handwritten in test booklet | Human scored: centrally scored by DRC |
| 2–3 | A | 2 | PL1–PL3 | Writing constructed response | Polytomous constructed response; handwritten in test booklet | Human scored: centrally scored by DRC |
| 2–3 | B/C | 2 | PL2–PL5 | Writing constructed response | Polytomous constructed response; handwritten in test booklet | Human scored: centrally scored by DRC |

| Grade-Level Cluster | Tier | Number of Tasks | Targeted PL Range | Task Type | Response Format | Scoring Procedures |
|---------------------|------|-----------------|-------------------|------------------------------|---|---------------------------------------|
| 4–5 | A | 2 | PL1–PL3 | Writing constructed response | Polytomous constructed response; handwritten in response booklet or keyboarded in test platform | Human scored: centrally scored by DRC |
| 4–5 | B/C | 2 | PL2–PL5 | Writing constructed response | Polytomous constructed response; handwritten in response booklet or keyboarded in test platform | Human scored: centrally scored by DRC |
| 6–8 | A | 2 | PL1–PL3 | Writing constructed response | Polytomous constructed response; handwritten in response booklet or keyboarded in test platform | Human scored: centrally scored by DRC |
| 6–8 | B/C | 2 | PL2–PL5 | Writing constructed response | Polytomous constructed response; handwritten in response booklet or keyboarded in test platform | Human scored: centrally scored by DRC |
| 9–12 | A | 2 | PL1–PL3 | Writing constructed response | Polytomous constructed response; handwritten in response booklet or keyboarded in test platform | Human scored: centrally scored by DRC |

As shown in Figure 18, the format of the Writing test is tiered. As Writing tasks are polytomous and elicit a range of student performances, each task is targeted to elicit language across a range of proficiency levels, rather than targeted to a single proficiency level. Tier A consists of tasks written to elicit language up to PL 3, while Tier B/C tasks are designed to elicit language up to PL 5. This is indicated by the large number in the colored rectangle in the figure. However, for both tiers of the test, DRC raters score students' responses to all tasks using the entire breadth of the scoring scale. Students can theoretically score anywhere from 0 to 9 on any task (in terms of the raw scores in the scoring scale), although the design of some tasks limits the possible scores. For example, Tier A tasks are not designed to elicit extended responses, so although the tasks are scored using the entire scale, these tasks do not elicit language above PL 4. Likewise, although Tier B/C tasks are designed to elicit extended discourse so that students can display proficiency at PL 5 or even PL 6, students' performances on these tasks may range from PL 1 to PL 6.

Figure 18.

Format of the Writing Test



Beginning with Series 501, both tiers consist of two tasks. Prior to Series 501, all test forms had three tasks, except for grade 1 Tier A, which consisted of four tasks. This change was made starting with Series 501 to accommodate an embedded field test design for field testing Series 502 Writing tasks. Tier A tasks target a single WIDA Standard; for all grade-level clusters except grade 1, Task I targets Language of Language Arts and Task II targets Language of Science, while for grade 1, Task I targets Language of Science and Task II targets Language of Language Arts. Tier B/C tasks integrate more than one WIDA Standard; Task I integrates Language of Language Arts and Language of Social Studies, and Task II integrates Language of Math and Language of Science. The ways in which the Standards are targeted by these tasks vary across grade levels and are spelled out in the generative item specifications.

The design of the embedded Writing field test for Series 602 is described in greater detail in Section 2.3.2.3.

Placement into tiers on the Writing test depends on the scores that students receive based on their performances on the Listening and Reading tests (which the test engine scores automatically). To determine how to best place each student into an appropriate tier, the CAL psychometrics team carried out logistic regression analyses to examine the relationship between student performance on the Listening and Reading tests administered in 2015–2016 and their performance on the Writing test. They then used this information to program an algorithm into the ACCESS Online test that the test engine uses to determine which tier of the Writing test to administer to each student. The purpose of the algorithm is to place students who are predicted to score above PL 3.0 into Tier B/C for the Writing test. All other students are placed into Tier A.

Although timing guidance is included in the test administrator manual, the Writing test is untimed.

2.2.4 Speaking

For the ACCESS Speaking test, Table 4 shows, for each grade-level cluster and tier, the number of tasks, the targeted range of WIDA proficiency levels, the task type, the response format, and the scoring procedure.

Table 4.
Number and Types of Tasks on the Speaking Test

| Grade-Level Cluster | Tier | Number of Tasks | Targeted PL Range | Task Type | Response Format | Scoring Procedures |
|---------------------|-------|-----------------|-------------------|-------------------------------|---------------------------------|---------------------------------------|
| 1 | Pre-A | 3 | PL1 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 1 | A | 6 | PL1-PL3 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 1 | B/C | 6 | PL3-PL5 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 2-3 | Pre-A | 3 | PL1 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 2-3 | A | 6 | PL1-PL3 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 2-3 | B/C | 6 | PL3-PL5 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 4-5 | Pre-A | 3 | PL1 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 4-5 | A | 6 | PL1-PL3 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 4-5 | B/C | 6 | PL3-PL5 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 6-8 | Pre-A | 3 | PL1 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 6-8 | A | 6 | PL1-PL3 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |

| Grade-Level Cluster | Tier | Number of Tasks | Targeted PL Range | Task Type | Response Format | Scoring Procedures |
|----------------------------|-------------|------------------------|--------------------------|-------------------------------|---------------------------------|---------------------------------------|
| 6–8 | B/C | 6 | PL3–PL5 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 9–12 | Pre-A | 3 | PL1 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 9–12 | A | 6 | PL1–PL3 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |
| 9–12 | B/C | 6 | PL3–PL5 | Speaking constructed response | Polytomous constructed response | Human scored; centrally scored by DRC |

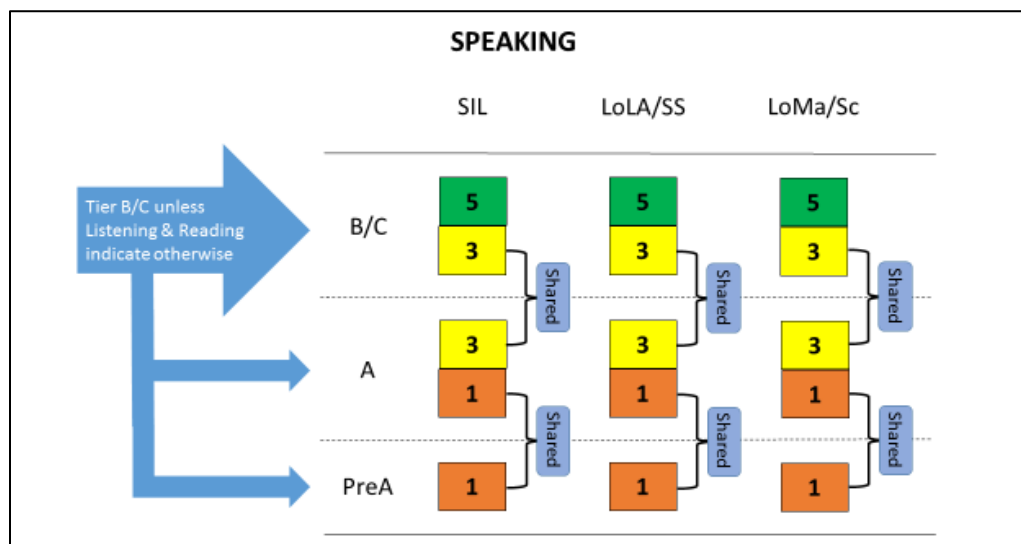
Figure 19 shows the format of the Speaking test. The Speaking test includes tasks that target language elicitation at three PLs: 1, 3, and 5. The tasks are grouped into thematic folders, each of which is aligned to one or two of the WIDA Standards. These folders are generally presented in the same order as the folders on the Listening and Reading tests; folders aligned to Social and Instructional Language are presented first, then folders aligned to Language of Language Arts, then folders aligned to Language of Math.

As shown in Figure 19, the Speaking test includes three tiers: Tier Pre-A, Tier A, and Tier B/C. Tier Pre-A includes tasks that target elicitation of language at PL 1. Tier A includes tasks that target elicitation of language at PLs 1 and 3. Tier B/C includes tasks that target elicitation of language at PLs 3 and 5.

A thematic panel refers to the folders across all tiers within a grade-level cluster that relate to a particular WIDA ELD Standard. In other words, the Tier B/C, Tier A, and Tier Pre-A folders that address Social and Instructional Language in each grade-level cluster make up a single thematic panel, with the PL 1 and PL 3 tasks shared across tiered folders in a panel. For example, within a Social and Instructional Language panel, the same PL 3 task appears on both the Tier A and the Tier B/C forms of the test, and the same PL 1 task appears on both the Tier Pre-A and Tier A forms of the test.

Figure 19.

Format of the Speaking Test



As with the Writing test, placement of students into the three tiers on the Speaking test depends on their performance on the Listening and Reading tests. Unlike Writing, the Speaking test has one additional tier, Tier Pre-A. Students are placed into Tier Pre-A when their scores on both the Listening and Reading tests are below PL 2.0. The Speaking Pre-A tier is designed to meet the needs of students in the very early stages of English language development. As noted previously, these tasks are targeted to the P1 level. DRC raters score students' responses to these tasks using a modified version of the full Speaking rating scale (see Section 4.2).

The process for placing students into Tiers A and B/C for the Writing test is analogous to the process used for tier placement for the Speaking test. The CAL psychometrics team carried out logistic regression analyses using test data for all students who were administered the assessment in 2015–2016 (i.e., the first year of the ACCESS Online assessment) to examine the relationship between students' performances on the Listening and Reading tests and their performance on the Speaking test. They used this information to program an algorithm into the ACCESS 2.0 Online test that the test engine used to determine which tier of the Speaking test to administer to each student. The purpose of the algorithm is to place students who are predicted to score above PL 3.0 into Tier B/C for the Speaking test, based on their performances in the Listening and Reading tests, and to place all other students into Tier A (except for those students who, as noted previously, are routed into Tier Pre-A).

Although timing guidance is included in the test administrator manual, the Speaking test is untimed.

2.3 Test Construction

2.3.1 Item and Task Development

The ACCESS item/task development process spans approximately three years and follows a standardized test development cycle. Each cycle begins with the development of a Refreshment Plan. The CAL TD team develops the Refreshment Plan, taking several factors into consideration, including empirical item/task performance, length of time that folders have been on the test, item/task-specification level information, and the success (or lack thereof) in refreshing the test for each targeted slot in the previous cycle. The CAL TD team presents the Refreshment Plan to the WIDA Assessment team for approval, with ultimate signoff by WIDA's Director of Test Development.

Upon receiving sign-off on the Refreshment Plan, the CAL TD team then determines which item/task specifications need to be updated or replaced and which can move forward as is. Generally, the CAL TD team updates or replaces item/task specifications for two reasons:

- The CAL TD team analyzes prior items and tasks that could not be used operationally due to fit issues, or in cases where the item or task fit was acceptable, an item or task difficulty measure that was outside of the range for the intended slot on the test. The purpose of this analysis is to determine if the poor performance is due to item/task mechanics (e.g., an issue with the wording of the passage or stem, a distractor that is too attractive) or if there is a deeper item/task-specification issue that cannot be resolved (e.g., the specification is difficult to operationalize successfully). In the latter case, the CAL TD team can update the specification (usually focused on updating the MPIs) or completely replace it, depending on the specific situation.
- The CAL TD team also updates or replaces item/task specifications as content standards change. As noted previously, the ACCESS item/task specifications include explicit connections to the content standards. If an update to the relevant content standard makes an ACCESS item/task specification obsolete, the CAL TD team revises or replaces the specification.

Once updates to item/task specifications are complete, item and task development begins. The generation of initial item/task content occurs in two interconnected steps. First, the CAL TD team initiates a process of theme generation. In the ACCESS item/task specifications, the CAL TD team writes each specification to a broad topic related to the given WIDA ELD Standard, and a theme is a more focused instantiation of the topic. For example, if the topic for a Language of Social Studies item/task specification for grades 4–5 is U.S. history, an example of an appropriate theme might be “the Industrial Revolution.”

The CAL TD team and WIDA TD staff are responsible for recruiting classroom English as a second language (ESL) and content teachers (from across the WIDA consortium) with experience teaching the academic content associated with one or more of the WIDA ELD Standards (including educators with experience working with English learners with disabilities), and the CAL TD team provides these educators with key parts of the item/task specification document (i.e., the topic, the MPIs, and guidelines for selecting a good theme). Then, the CAL

TD team asks educators to propose themes related to the topic, along with possible directions for each item or task, which are grade-level appropriate. After the theme generation process is complete, the CAL TD team reviews the list of themes to identify those that will become the focus of item/task writing. This determination is based on several factors, including operationalizability on a large-scale assessment (since many ideas from educators are well suited for the classroom but do not clearly translate to the assessment context), themes currently in use on the assessment, and bias and sensitivity considerations.

The team then assigns themes to professional item/task writers to develop the initial item/task content. The team recruits individuals with prior experience developing ESL or English language arts items/tasks, preferably in the context of large-scale, standardized assessments, but individuals with other experience (such as experience writing items/tasks for language tests in languages other than English, and experience with English placement tests for the college/university setting) are also considered. These item writers do not need to be educators in WIDA states; experience with item writing for ESL/ELA assessments is the primary criterion for inclusion. All item/task writers, both new item/task writers and those returning from the previous test development cycle, participate in an introductory training, and the team provides them with extensive documentation regarding writing items/tasks for ACCESS, including an Item Writing Handbook and ancillary documents (i.e., checklists, item/task specifications, templates) to complete their assignments. One or more CAL language testing specialists work with each item/task writer, providing feedback on the item content.

After item/task writing is complete, CAL language testing specialists and test development managers review the folders, using a standard checklist, to determine which folders will undergo further development and which will be retired. Folders then go to their first external review, the Standards Expert review.

During the Standards Expert review, educators provide feedback about the overall grade-level appropriateness of the language and content of the items/tasks to ensure that no drift, in terms of grade-level appropriateness of the content or the language, has occurred between initial theme generation and item/task writing. The CAL TD team and WIDA TD staff are jointly responsible for recruiting educators with ESL and content-area expertise from across the WIDA Consortium to serve as Standards Experts. CAL language testing specialists prepare a short questionnaire with both yes/no and open-ended questions about each folder and send the questionnaires and folders to the Standards Experts.

Subsequent to the Standards Expert review, all content proceeds through a rigorous folder refinement stage internal to CAL. Folder refinement includes numerous steps, including additional research and sourcing/fact-checking, meticulous review against a comprehensive, industry-standard item/task development checklist with peer review that other language testing specialists carry out, as well as review by test development managers and the Director of Test Development and successive rounds of revision before sign-off. During this stage, all aspects of the items and tasks are scrutinized: the WIDA proficiency level of the stimulus, the graphic support, the question stems, and response options (for the Listening and Reading tests) and task prompts (for the Speaking and Writing tests). The CAL TD team also conducts mock administrations. During this phase, CAL language testing specialists produce other

ancillary materials, such as Test Administrator Scripts. Upon sign-off, the CAL TD team works with the CAL Production and Tech teams to generate the graphics used on the test and to begin the development of the question and test interoperability (QTI) packages for the online assessment. A QTI package is a collection of files that contain all the item/task content, including assets such as graphics and audio files, coded so that the test engine can read them. There is one QTI package for each folder on ACCESS. Once the graphics are generated, CAL language testing specialists inserted them into the folders and conducted layout review and fact-checking (with test development manager sign-off) to ensure that the items and tasks are ready for external Content Review and Bias and Sensitivity Review.

Content Review and Bias and Sensitivity Review are external reviews that educators and WIDA TD staff carry out on ACCESS items and tasks. WIDA TD staff are responsible for assembling these panels by recruiting educators of multilingual learners from around the consortium, including culturally, racially, and linguistically diverse educators who reflect the population of students that take WIDA assessments. WIDA employs several criteria when recruiting educators to perform these tasks. The criteria used to recruit educators to conduct Content Reviews differ somewhat from the criteria used to recruit educators to conduct Bias and Sensitivity Reviews.

Educators conduct Content Reviews by grade-level cluster (G1, G2–3, G4–5, G6–8). The educators who are recruited to review a particular grade-level cluster's content (four reviewers per grade-level cluster) have experience teaching English learners and are either currently teaching students who are in that grade-level cluster or have extensive prior experience teaching students who are in that grade-level cluster. Additionally, educators serving on each panel represent different content areas. WIDA TD staff seek to ensure that each panel includes at least one educator who has teaching experience in each of the following content areas: ELA, Science, Math, Social Studies, and Special Education. Additionally, during the recruitment process, WIDA TD staff seek to ensure diversity and balance across (1) consortium states, (2) school locale (rural/suburban/urban), and (3) years of teaching experience. The CAL TD team and WIDA TD staff first train the Content Review Panel on the procedures and scope of the review. The panelists are introduced to the test layout, instructed on the logistics of the review, and trained to use the review checklist. The panel members then individually review each item and task, followed by a collective discussion of each item and task to determine (1) whether the content is accessible and relevant to students in the targeted grade-level cluster, (2) is at the targeted WIDA proficiency level, and (3) matches the Model Performance Indicator from the WIDA English Language Development Standards that it is intended to assess.

The Bias and Sensitivity Review Panel ensures that test items and tasks are free of material that (1) might favor any subgroup of students over another on the basis on gender, race/ethnicity, home language, religion, culture, region, or socioeconomic status, and (2) might be upsetting to students. Educators conduct Bias and Sensitivity Reviews by grade groupings (e.g., G1–3, G4–5, G6–8, and G9–12). The educators who are recruited to review a particular grade-level cluster's content (5 or 6 reviewers per grade grouping) are educators or school administrators who have experience teaching English language learners and are either currently teaching students who are in that grade-level cluster or have extensive prior experience teaching students who are in that grade-level cluster. WIDA TD staff employ additional criteria to ensure

that a variety of perspectives are represented on each panel. These criteria include recruiting at least one educator with experience in Special Education to serve on each panel. Additionally, during the recruitment process, WIDA TD staff seek to ensure diversity and balance across (1) consortium states, (2) school locale (rural/suburban/urban), and (3) years of teaching experience. The CAL TD team and WIDA TD staff conduct training for all new and returning reviewers before any items or tasks are reviewed. The panel members then individually review each item and task, followed by a collective discussion of each item and task to determine if any bias or sensitive topics are detected in the items/tasks, and if so, what the CAL TD team can do to remediate the issues. The CAL TD team and WIDA TD staff facilitate the reviews and take extensive notes to capture all feedback during the reviews. WIDA TD staff also conduct a separate, asynchronous review around the time of the Content Review and Bias and Sensitivity Review, using the same materials that the educators review, and provide written feedback on the materials.

The CAL language testing specialists compile all the Content Review and Bias and Sensitivity Review feedback from educators and from WIDA TD staff, and then work to implement the feedback, with the CAL test development manager sign-off as a final step. The CAL Test Production and Tech teams then revise the graphics and the QTI packages. The input and feedback from educators at various stages in the item/task development process serves as evidence that each item and task is appropriate for the age and grade-level cluster for which it is intended.

Tasks in the Writing domain undergo one additional step: a small-scale tryout with educators and students. Given the changes to the Writing test over the past few years, including a change from three to two operational tasks, along with changes to task specifications to better align the Writing tasks with classroom practice, these tryouts allow the CAL TD team to evaluate whether each Writing task will effectively elicit language at its targeted WIDA proficiency level. For the Writing tryouts, the CAL TD team and WIDA TD staff jointly recruit educators with appropriate numbers of students at the targeted proficiency levels (approximately 15 students per task) to participate. The CAL test development manager for Writing prepares a recruitment flyer, which the CAL TD team and WIDA TD staff circulate to educators.

The CAL TD team circulates the flyer to educators who have previously participated in the tryouts, and WIDA TD staff circulate the flyer through WIDA's regular SEA/LEA communications emails. Due to the small-scale nature of the tryouts, the recruitment is ultimately a convenience sample, although CAL and WIDA strive to obtain a sample with geographic diversity (i.e., educators distributed throughout the consortium, with a mix of urban, suburban, and rural representation). In addition, the tasks target different proficiency levels/tiers at the different grade level clusters, so recruiting educators with students at these targeted proficiency levels and grade level clusters is another requirement. Finally, we do not recruit educators who have already reviewed the tasks in development during Bias and Sensitivity and Content review to participate in tryouts. If the CAL TD team determines that the first round of recruitment has failed to find educators with students at the appropriate proficiency levels for all grade-level clusters and tiers, the CAL test development manager for Writing identifies the grade-level clusters and proficiency levels/tiers with gaps and provides this information to the WIDA TD staff, who can then do targeted recruitment based on existing

databases of educators who have indicated willingness to participate in test development activities.

The educators administer the tasks to their students and send the students' written responses back to CAL for analysis. The students and the educators also fill out short surveys about the tasks. The students each fill out a six-question/prompt survey answering questions like "I understood what to do." and "This is an interesting topic to write about." The educators complete an eight-question survey focusing on the effectiveness and appropriateness of the task input and graphics, the comparability of the task to first-draft writing in class, and student familiarity and engagement with the task content.

CAL language testing specialists conduct qualitative analyses of the student responses and the survey data and use the results to inform any final revisions to the tasks prior to field testing. For some tiers, the tryouts also inform which task moves on to field testing and which is postponed, in cases where only a single task is field tested. (See Section 2.3.2 for more information regarding the field test design.)

After the CAL language testing specialists complete edits from the Content Review and Bias and Sensitivity Review (and tryout edits for Writing), they then prepare the folders for final production. Additionally, they produce audio recording scripts for professional audio recording, arrange for recording the audio files, complete extensive quality control checks for both content and technical specifications of the audio (e.g., file types, recording quality, and compression levels), conduct final layout reviews, and perform key checks for the Listening and Reading tests. Both the CAL TD team and WIDA TD staff conduct quality control checks of the QTI. The WIDA TD staff sign off on all materials before DRC builds the final test forms in the test engine. Items and tasks that reach this point then go through field testing processes, described in the next subsection by domain.

Throughout the item/task development process, the CAL TD team focuses on issues of fairness. The team applies the seven Universal Design of Assessment (UDA) principles when creating items and tasks:

- Inclusive assessment population
- Precisely defined constructs
- Accessible, nonbiased items/tasks
- Amenability to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

Additionally, when CAL's TD managers, WIDA TD staff, and external reviewers conduct Standards Expert reviews, Content Reviews, and Bias and Sensitivity Reviews, they use checklists that ask them to consider the seven principles of universal design as they are reviewing each item and task.

In recent years, WIDA has placed additional focus on ensuring that the items and tasks, and especially the graphics, are amenable to accommodations by involving WIDA's Accessibility and Accommodations Team directly in the item/task review process. WIDA's Accessibility and

Accommodations Team helped CAL's TD team develop principles for graphics development and for eliminating language that is biased towards students with sight, and WIDA's Accessibility and Accommodations Team also reviews the items and tasks during development to help CAL identify areas that still need to be addressed.

Through maintaining a focus on fairness throughout the test development cycle by integrating the principles of UDA in various steps, the CAL TD team strives to ensure that ACCESS Online items and tasks are best positioned to be maximally fair for all populations.

2.3.2 Field Testing

2.3.2.1 Listening

DRC field tested the Listening items developed for Series 602 as embedded folders during the operational administration of Series 601. The embedded field test folders contained items that featured innovative formats, including hot spot items (i.e., the student clicks on an area of the screen to respond) and drag-and-drop items (i.e., the student drags an image/text to a specified screen area to respond).

For Series 602, DRC field-tested a total of 87 Listening items (29 folders), across all five grade-level clusters, as indicated in Table 5.

Table 5.**Number of Field Test Folders and Items for the Series 602 Listening Test**

| Grade-Level Cluster | Tier Pool | Number of Folders to Refresh | Number of Overage Folders | Total Number of Field Test Folders | Total Number of Field Test Items | Standards Addressed in FT |
|----------------------------|-----------------------|-------------------------------------|----------------------------------|---|---|----------------------------------|
| 1 | Entry | 1 | 0 | 1 | 3 | SI |
| 1 | A | 1 | 0 | 1 | 3 | SS |
| 1 | B | 2 | 0 | 2 | 6 | MA, SS |
| 1 | C | 0 | 0 | 0 | 0 | Not Applicable |
| 2-3 | Entry | 1 | 0 | 1 | 3 | SI |
| 2-3 | A | 2 | 0 | 2 | 6 | LA, SS |
| 2-3 | B | 4 | 0 | 4 | 12 | MA, SS, SC, LA |
| 2-3 | C | 0 | 0 | 0 | 0 | Not Applicable |
| 4-5 | Entry | 0 | 0 | 0 | 0 | Not Applicable |
| 4-5 | A | 1 | 0 | 1 | 3 | SS |
| 4-5 | B | 2 | 0 | 2 | 6 | LA |
| 4-5 | C | 3 | 0 | 3 | 9 | LA, SS |
| 6-8 | Entry | 1 | 0 | 1 | 3 | SI |
| 6-8 | A | 1 | 0 | 1 | 3 | SS |
| 6-8 | B | 3 | 0 | 3 | 9 | SS, SC, LA |
| 6-8 | C | 1 | 0 | 1 | 3 | SS |
| 9-12 | Entry | 1 | 1 | 2 | 6 | SI |
| 9-12 | A | 1 | 0 | 1 | 3 | SC |
| 9-12 | B | 3 | 0 | 3 | 9 | MA, SC, LA |
| 9-12 | C | 0 | 0 | 0 | 0 | Not Applicable |
| Total | Not Applicable | 28 | 1 | 29 | 87 | Not Applicable |

Each student received one Listening field test folder embedded in the operational test. Field test folders are targeted to refresh a specific operational folder on the test, and field test folder specifications include the stage, WIDA ELD Standard, and tier pool target (i.e., Entry, A, B, or C) of the folder. Students received the embedded field test folder at the stage targeted for refreshment, with administration randomized so that half of the students saw the field test folder before the corresponding operational folder, and half saw the operational folder before the field test folder. Field test folders were administered to those students who were routed to take the operational folder that was either at the same tier or adjacent to the tier that the field test folder targeted.

When DRC drew the field test samples, 50% of the sample were students who were routed to the tier that the field test folder targeted, and the other 50% were students who were routed to adjacent tiers. (If there were adjacent tiers both above and below the field test target, then 25%

of the sample were students routed to each of those tiers.) In cases where the field test folder was to be placed in one of the entry stages, students receiving that field test folder took it directly after the pair of operational entry folders. CAL set the field test sample targets for the Listening test at a minimum of 3,000 responses per folder.

Because CAL's psychometrics team used the Listening field test data in the pre-equating analysis, their sample size requirement of 3,000 was much higher than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, to ensure that the pre-equated parameter estimates would be stable. Linacre (1994), citing Wright and Douglas's (1975) formulation, explained how to determine the minimum sample required for calibrating dichotomous items to achieve various levels of estimation precision and confidence intervals. With a sample size of 3,000, one can be 95% confident that no item parameter will be more than ± 0.1 logit away from its true value. The sample sizes for all field test folders exceeded the minimum requirement of 3,000, except for one Listening grade-level cluster 4–5 Tier A folder, which had a sample size of 2,800, due to the fact that the population of grades 4–5 tier A students comprise a smaller portion of the population than students in other grade-level clusters and tiers.

After CAL's psychometrics team accessed the field test data, they analyzed students' responses to the items in the field test folders to determine each item's psychometric properties, and folders for which all three items met established psychometric standards (as described below) were eligible for inclusion in the next year's operational test.

The team then classified each item using the classification system shown in Table 6. If all three items in a folder were green, the entire folder was eligible for operational use. If one or more items were red, the folder was no longer considered appropriate for operational use. If one or more items were yellow, the Post-Field Test Review Panel reviewed the content of each item, along with relevant statistics contained in the distractor analyses (e.g., the mean ability of students selecting the key vs. the mean ability of students selecting the distractors; infit and outfit statistics for each response option, the point measure correlation of each response option, the percentage of students selecting each response option), to determine if each item would be reclassified as green or red. If all yellow items in a folder were reclassified as green (and there were no red items in the folder), the folder was deemed appropriate for operational testing.

Table 6.

CAL's Post-Field Test Review Classification System for Series 602

| Color | Interpretation | Definition |
|--------|---|---|
| Green | Appropriate for operational testing | A- or B-level DIF AND a p value $\geq .85$ OR infit and outfit ≤ 1.20 |
| Yellow | Content review is required to confirm item is appropriate for operational testing | C-level DIF OR infit/outfit > 1.20 and ≤ 1.50 Three-response choice item with p value $\leq .40$ and outfit < 1.75 Four-response choice item with p value $\leq .35$ and outfit < 1.75 |
| Red | Not appropriate for operational testing | Infit/outfit > 1.50 |

2.3.2.2 Reading

DRC field-tested the Reading items developed for Series 602 as embedded items during the operational administration of Series 601. The embedded field test folders contained items that featured innovative formats, including hot spot items (i.e., the student clicks on an area of the screen to respond) but no drag-and-drop items.

For Series 602, DRC field-tested a total of 111 Reading items (37 folders), across all five grade-level clusters, as indicated in Table 7.

Table 7.**Number of Field Test Folders and Items for the Series 602 Reading Field Test**

| Grade-Level Cluster | Tier Pool | Number of Folders to Refresh | Number of Overage Folders | Total Number of Field Test Folders | Total Number of Field Test Items | Standards Addressed in FT |
|----------------------------|-----------------------|-------------------------------------|----------------------------------|---|---|----------------------------------|
| 1 | Entry | 1 | 0 | 1 | 3 | SI |
| 1 | A | 3 | 0 | 3 | 9 | SS, LA, MA |
| 1 | B | 0 | 0 | 0 | 0 | Not Applicable |
| 1 | C | 0 | 0 | 0 | 0 | Not Applicable |
| 2-3 | Entry | 1 | 0 | 1 | 3 | SI |
| 2-3 | A | 1 | 0 | 1 | 3 | SC |
| 2-3 | B | 2 | 0 | 2 | 6 | MA, SC |
| 2-3 | C | 2 | 0 | 2 | 6 | SS, SC |
| 4-5 | Entry | 2 | 1 | 2 | 6 | SI |
| 4-5 | A | 2 | 0 | 2 | 6 | LA, MA |
| 4-5 | B | 4 | 1 | 5 | 15 | LA, MA, SS |
| 4-5 | C | 1 | 0 | 1 | 3 | SC |
| 6-8 | Entry | 1 | 1 | 2 | 6 | SI |
| 6-8 | A | 1 | 0 | 1 | 3 | LA |
| 6-8 | B | 3 | 2 | 5 | 15 | MA, SS |
| 6-8 | C | 2 | 0 | 2 | 6 | SS, LA |
| 9-12 | Entry | 1 | 1 | 2 | 6 | SI |
| 9-12 | A | 2 | 0 | 2 | 6 | LA, MA |
| 9-12 | B | 3 | 0 | 3 | 9 | LA, SC |
| 9-12 | C | 0 | 0 | 0 | 0 | Not Applicable |
| Total | Not Applicable | 31 | 6 | 37 | 111 | Not Applicable |

DRC administered the embedded Reading field test in the same way as the embedded Listening field test. As with the Listening test, CAL set the field test sample targets for the Reading test at a minimum of 3,000 responses per folder. The sample sizes for all field test folders exceeded the minimum requirement of 3,000.

After CAL's psychometrics team accessed the field test data, they analyzed students' responses to the items in the field test folders to determine each item's psychometric properties, and folders for which all three items met established psychometric standards (as described in Section 2.3.2.1) were eligible for inclusion in the next year's operational test.

2.3.2.3 Writing

DRC administered the Series 602 Writing tasks in an embedded field test model. For Series 602, a total of 10 Writing tasks were field-tested, as indicated in Table 8.

Table 8.

Number of Field Test Tasks for Series 602 Writing

| Grade-Level Cluster | Tier | Number of Folders to Refresh | Number of Folders Field-Tested | Standards Addressed in FT |
|----------------------------|-----------------------|-------------------------------------|---------------------------------------|----------------------------------|
| 1 | A | 1 | 1 | LA |
| 1 | BC | 1 | 1 | LA/SS |
| 2–3 | A | 1 | 1 | LA |
| 2–3 | BC | 1 | 1 | LA/SS |
| 4–5 | A | 1 | 1 | LA |
| 4–5 | BC | 1 | 1 | LA/SS |
| 6–8 | A | 1 | 1 | LA |
| 6–8 | BC | 1 | 1 | LA/SS |
| 9–12 | A | 1 | 1 | LA |
| 9–12 | BC | 1 | 1 | LA/SS |
| Total | Not Applicable | 10 | 10 | Not Applicable |

All students received a field test folder that was appended to their operational assessment. Students received a field test folder in the tier that corresponded to their operational tier. CAL targeted a sample of 500 students per task. This was much higher than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, making it likely that, for each of the nine scale categories, there would be at least 10 students whose responses to the task would warrant receiving scores in that category, as Linacre (2002a) recommended for polytomously scored tasks. If raters assign fewer than 10 scores in each scale category, then the category statistics for that category tend to be unstable. Historically, the distribution of scores that raters assign to students' responses to the Writing tasks tends to be highly concentrated in the middle of the score distribution (i.e., exhibit a central tendency effect), with raters assigning relatively fewer scores in the categories at the high end of the score scale. Therefore, CAL targeted a sample size of 500 to ensure that there would be students for analysis whose responses to the task would warrant receiving scores at the high end of the 9-category score scale. Use of this larger sample size also provided examples of students' responses that received scores in the higher scale categories that trainers could use as anchor papers for rater training.

DRC administered the field test under standard testing conditions. The field test used the online interface with keyboarded responses for grades 4–12 and paper booklets with handwritten responses for grades 1–3. For the Writing field test, DRC raters scored the

students' responses to the field test tasks. DRC performed a 20% read-behind as a quality control measure, with the first score assigned as the score of record.⁴

2.3.2.4 Speaking

All Tier A and B/C students received a Speaking field test folder that was appended to their operational Speaking assessment. Tier Pre-A was not included in the field test. DRC field tested a total of 30 folders (15 panels) for Series 602, with a target sample size of 500 students per folder. This is much higher than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, making it likely that, for each of the four scale categories, there would be at least 10 students whose responses to the task would warrant receiving scores in that category, as Linacre (2002a) recommended for polytomously scored tasks. Historically, the distribution of scores that raters assign to students' responses to the Speaking tasks tends to be highly concentrated in the middle of the score distribution (i.e., exhibit a central tendency effect), with raters assigning relatively fewer scores in the categories at the high end of the score scale. Therefore, CAL targeted a sample size of 500 to ensure that there would be students for analysis whose responses to the task would warrant receiving scores at the high end of the 4-category score scale.⁵ Use of this larger sample size also provided examples of students' responses that received scores in the higher scale categories that trainers could then use as anchor papers for rater training.

DRC-trained raters scored students' responses to the field test Speaking tasks, with a 20% read-behind as a quality control measure and the first score as the score of record.

Students received a Speaking field test folder in the tier that corresponded to their operational tier. For Series 602, CAL field tested a total of 28 Speaking folders, as indicated in Table 9.

⁴ The purpose of the 20% read-behind is to monitor rater performance on a daily basis. (See Section 4.2 below). If the read-behinds detect that one rater is consistently scoring inaccurately, DRC can rescore all of the students' responses to tasks scored by that rater, and the rater can be retrained or terminated. Ratets go through significant training and qualification prior to live scoring, and they are monitored daily through validity and recalibration tasks, so a scenario where a rater is consistently anomalous in his or her ratings would be uncommon, and it would be detected and corrected immediately.

⁵ Technically, the score scale includes 5 categories, including "No Response (in English)."

Table 9.**Number of Field Test Tasks for Series 602 Speaking**

| Grade-Level Cluster | Tier | Number of Folders to Refresh | Number of Folders Field Tested | Standards Addressed in FT |
|----------------------------|-----------------------|-------------------------------------|---------------------------------------|----------------------------------|
| 1 | A | 2 | 3 | SIL, MA/SC |
| 1 | BC | 2 | 3 | SIL, MA/SC |
| 2–3 | A | 2 | 3 | SIL, MA/SC |
| 2–3 | BC | 2 | 3 | SIL, MA/SC |
| 4–5 | A | 2 | 3 | SIL, MA/SC |
| 4–5 | BC | 2 | 3 | SIL, MA/SC |
| 6–8 | A | 2 | 3 | SIL, MA/SC |
| 6–8 | BC | 2 | 3 | SIL, MA/SC |
| 9–12 | A | 2 | 2 | MA/SC |
| 9–12 | BC | 2 | 2 | MA/SC |
| Total | Not Applicable | 20 | 28 | Not Applicable |

2.3.3 Item/Task Review and Selection

After the analysis of field test data, a panel consisting of members of the WIDA TD and psychometrics staff, the CAL TD Team, and the CAL psychometrics team conducted an item/task selection meeting to determine which of the field-tested folders would be placed on the Series 602 operational assessment. Results from qualitative and quantitative analyses guided the selection of operational items and tasks.

In the domains of Listening and Reading, item selection was a two-step process. First, the Item Selection Panel reviewed the field test results. CAL's psychometrics team used a three-tier color-coding system for field test review. Items are coded as "green," "yellow," or "red," and CAL's psychometrics team then assigned each folder a color based on the least favorable item in the folder. In other words, a folder with a red item was always coded as red, a folder with a yellow item (but no red items) was coded yellow, and folders were coded green only when all items were green.

Items were coded by color according to the following criteria:

- If an item showed C-level or CC-level differential item functioning (DIF), it was automatically coded yellow. Any items that showed this level of DIF were subject to an extra round of review (to determine if anything in the item could be detected that clearly indicates bias) prior to item selection (see Part 2, Section 2.2 for further detail). The CAL psychometrics team provided the Item Selection Panel with the report of the DIF review.
- Items were coded as green if they had infit and outfit values ≤ 1.20 . As outfit and infit values are sensitive to students' unexpected responses to items that are very easy for them, any item with a p value > 0.85 was automatically coded as green, even if it had fit values outside of these thresholds.
- Items with infit and outfit values > 1.20 and < 1.50 were coded as yellow. As outfit values are also sensitive to students' unexpected responses to items that are very hard for

them, items with p values close to chance (0.40 for a three-response item and 0.35 for a four-response item) were coded as yellow if outfit was >1.20 and <1.75 .

- Items that did not meet these criteria were coded as red.

The task of the Item Selection Panel in this first stage was to review all yellow folders and recode them as “green,” meaning “appropriate for operational use,” or “red,” meaning “not appropriate for operational use.” The panel reviewed the content of each yellow item, along with relevant statistics like those derived from distractor analyses (e.g., the mean ability of students selecting the key vs. the mean ability of students selecting the distractors; infit and outfit statistics for each response option, the point measure correlation of each response option, the percentage of students selecting each response option), to determine if the item would be reclassified as green or red. If all yellow items in a folder were reclassified as green (and there were no red items in the folder), the folder was deemed appropriate for operational testing.

In the next stage, the set of green folders, which the panel had deemed appropriate for operational use, became the pool of folders for item selection. The panelists selected folders (through a process of discussion and consensus building) with attention to the difficulty of each item within a folder, the mean item difficulty of the items within a folder, and the content of a folder.

Table 10 and Table 11 provide the numbers of continuing and new items per grade-level cluster for the Listening and Reading tests. For further detail on item statistics, including a summary of the number of items used as anchors across years, see Part 2, Sections 2.1 and 2.7.

Table 10.

Number of New and Continuing Items on ACCESS Online, Series 602 Listening Test, by Grade-Level Cluster

| Grade-Level Cluster | Number of New Items | Number of Continuing Items | Total Number of Items |
|---------------------|---------------------|----------------------------|-----------------------|
| 1 | 3 | 51 | 54 |
| 2-3 | 6 | 48 | 54 |
| 4-5 | 6 | 48 | 54 |
| 6-8 | 12 | 42 | 54 |
| 9-12 | 9 | 45 | 54 |

Table 11.**Number of New and Continuing Items on ACCESS Online, Series 602 Reading Test, by Grade-Level Cluster**

| Grade-Level Cluster | Number of New Items | Number of Continuing Items | Total Number of Items |
|----------------------------|----------------------------|-----------------------------------|------------------------------|
| 1 | 6 | 66 | 72 |
| 2-3 | 9 | 63 | 72 |
| 4-5 | 9 | 63 | 72 |
| 6-8 | 18 | 54 | 72 |
| 9-12 | 9 | 63 | 72 |

In the domains of Writing and Speaking, the Task Selection Panel considered results from both qualitative and quantitative analyses of the students' responses to the tasks. The CAL TD team reviewed student responses and DRC raters' comments on each of the field-tested tasks. They then integrated those observations with task statistics, including fit statistics, raw score distributions, and rater agreement indices. Based on the information they compiled, the team made a recommendation to present to the panel for each task. If the panel needed to choose between two tasks, the team identified the task that was most appropriate to place on the operational test, based on the evidence they had compiled. Alternatively, the team could recommend that the slot remain unrefreshed. In cases where there was only a single task, the team determined whether the associated evidence was sufficient to support placing the task on the operational test or whether that slot should remain unrefreshed.

Although the CAL TD team considered rater agreement indices and fit statistics when making their recommendations, they based those recommendations primarily on the results from their analyses of the following: (1) the qualitative data (i.e., whether students could successfully score in the intended range, and/or whether DRC raters observed major anomalies that could indicate that a given task was not performing as intended), (2) the raw score distributions of students' task performance, and (3) the task difficulty measures. The field-test tasks and the operational tasks that were tagged for refreshment should have comparable raw score distributions and task difficulty measures, they reasoned. The CAL TD team took this approach to ensure that the vertical scale was maintained from year to year. The panel then reviewed each recommendation and associated evidence and either accepted or rejected the recommendation; recommendations were generally rejected if the task difficulty measures and the raw score distributions of the field-test tasks varied too much from those of the operational tasks.

Table 12 and Table 13 provide the numbers of continuing and new tasks, per grade-level cluster, for the Writing and Speaking tests. For further detail on task statistics, including a summary of the number of tasks used as anchors across years, see Part 2, Sections 2.1 and 2.7.

Table 12.**Number of New and Continuing Tasks on ACCESS Online Series 602 Writing Test, by Grade-Level Cluster**

| Grade-Level Cluster | Tier | Number of New Items | Number of Continuing Items | Total Number of Items |
|----------------------------|-------------|----------------------------|-----------------------------------|------------------------------|
| 1 | A | 1 | 1 | 2 |
| 1 | B/C | 1 | 1 | 2 |
| 2-3 | A | 1 | 1 | 2 |
| 2-3 | B/C | 1 | 1 | 2 |
| 4-5 | A | 1 | 1 | 2 |
| 4-5 | B/C | 1 | 1 | 2 |
| 6-8 | A | 1 | 1 | 2 |
| 6-8 | B/C | 1 | 1 | 2 |
| 9-12 | A | 1 | 1 | 2 |
| 9-12 | B/C | 0 | 2 | 2 |

Table 13.**Number of New and Continuing Tasks on ACCESS Online Series 602 Speaking Test, by Grade-Level Cluster**

| Grade-Level Cluster | Tier | Number of New Tasks | Number of Continuing Tasks | Total Number of Tasks |
|----------------------------|-------------|----------------------------|-----------------------------------|------------------------------|
| 1 | Pre-A | 2 | 1 | 3 |
| 1 | A | 4 | 2 | 6 |
| 1 | B/C | 4 | 2 | 6 |
| 2-3 | Pre-A | 2 | 1 | 3 |
| 2-3 | A | 4 | 2 | 6 |
| 2-3 | B/C | 4 | 2 | 6 |
| 4-5 | Pre-A | 2 | 1 | 3 |
| 4-5 | A | 4 | 2 | 6 |
| 4-5 | B/C | 4 | 2 | 6 |
| 6-8 | Pre-A | 1 | 2 | 3 |
| 6-8 | A | 2 | 4 | 6 |
| 6-8 | B/C | 2 | 4 | 6 |
| 9-12 | Pre-A | 2 | 1 | 3 |
| 9-12 | A | 4 | 2 | 6 |
| 9-12 | B/C | 4 | 2 | 6 |

3. Test Administration

3.1 *Test Delivery*

ACCESS Online is typically administered between December and April of the academic year, with testing windows determined at the state level. The Reading and Listening tests are administered first (in either order), followed by Writing and Speaking (in either order). The test may be administered in several sessions within a single day or over a series of days.

3.1.1 Listening and Reading

Listening and Reading are the first domains assessed. Students may take these in either order. Students sit at individual computer monitors and take the Listening and Reading tests online. They use headsets to listen to directions for the Listening and Reading tests, as well as listen to the Listening items. Students use the computer interface to select their answers. Once a student selects an answer and clicks the Next button, the answer is final, and the student is not permitted to go back and change an answer. The Listening and Reading tests are untimed, but approximate administration times are provided in the following sections.

3.1.2 Writing

Students in grades 1–3 perform the Writing tasks on paper and handwrite their response.

Students in grades 4–12 perform the Writing tasks online. A student may provide handwritten or keyboarded responses, with the choice dependent on a combination of local, state, and consortium-wide policies, as follows:

- Grades 4–5: A decision is made at the local or state level as to whether handwriting or keyboarding is the default response mode. In districts where keyboarding is the default, the option exists to use handwriting as an accommodation.
- Grades 6–12: Keyboarding is the default, with the option to use handwriting as an accommodation.

3.1.3 Speaking

Speaking tasks are delivered online. Students listen to prompts via headsets that are equipped with microphones to capture their responses. The student receives extensive support via illustrations and multimodal (text and audio) input designed to provide sufficient context for the response, as well as a model student response that provides guidance on the level of linguistic complexity required to respond adequately (see Section 2.2.4).

3.2 Operational Administration

Before, during, and after a state's testing window, there are various roles that educators hold to ensure all tasks are carried out for successful test administration. These roles include test coordinators at the district and school level, test administrators, and, for online administration, technology coordinators. The test administrator administers and monitors the test and is responsible for managing student data prior to, during, and after testing. The test administrator manual and the District and School Test Coordinator Manual contain more information related to responsibilities and required training for the various roles. These manuals can be found on the [WIDA Secure Portal](#).

The training course within the WIDA Secure Portal where educators can access both training to become certified to administer ACCESS for ELLs as well as additional materials and resources to assist administrators and coordinators before, during, and after a state's testing window. Training courses include test preparation and administration tutorials and online administration quiz.

It cannot be understated that the roles of test administrator and technology coordinator are critical for the proper administration of the assessments. Proper training and familiarity with ACCESS for ELLs administration requirements is key to the validity of the test and the appropriate interpretations of ACCESS for ELLs test scores.

3.2.1 Administering the Test Practice

A test practice experience is provided to each student immediately prior to the administration of the individual test domain. The test practice acclimates the student to the test interface and the types of items the student may experience in the test. The test practice takes approximately 5 to 10 minutes, depending on how many questions students have about the directions or practice items. Additional time should be scheduled for students to go through the test practice again if needed. The narration within the test practice is included both as spoken audio and as text captioning displayed directly on the screen, allowing the student to be able to read along as the script is read aloud.

The test practice for each domain and grade-level cluster are available as stand-alone materials on the WIDA website (<https://wida.wisc.edu/assess/access/preparing-students/practice>) to help educators prepare students to take ACCESS for ELLs. Before each domain test of ACCESS for ELLs, each student is required to take the test practice for that domain. No data are collected regarding the test practice; these items/tasks are presented to the students specifically to help ensure that they understand how to navigate the test interface.

3.2.2 Listening Test Administration

The Listening test (including test practice items) is designed to take approximately 30 to 40 minutes. Students in all grades view the Listening prompts on the desktop, laptop, or tablet. Note that the approximate test administration time does not include convening students, taking attendance, or explaining test directions.

All Listening items are forced choices; in other words, students must respond to an item before they can proceed to the next item. In addition, once the students proceed to the next screen, they cannot return to any previous screens.

3.2.3 Reading Test Administration

The Reading test (including directions and practice items) is designed to take approximately 35 minutes. Students in all grades view the Reading prompts on the desktop, laptop, or tablet. Note that the approximate test administration time does not include convening students, taking attendance, or explaining test directions.

All Reading items are forced choices; in other words, students must respond to an item before they can proceed to the next item. In addition, once the students proceed to the next screen, they cannot return to any previous screens.

3.2.4 Writing Test Administration

All students in grades 1–3 complete the ACCESS for ELLs Writing test on paper. The test is group administered. For grades 6–12, all students view the Writing prompts on the desktop, laptop, or tablet. The default response mode is keyboarding. For grades 4–5, all students also view the Writing prompts on the device. However, each state determines whether the default response mode for students in grades 4–5 will be keyboarding or handwriting. If keyboarding is the default response mode, and upon logging in and starting the test a student expresses discomfort, concern, or anxiety about keyboarding, administrators may switch the student to responding to the Writing test on paper. For grades 6–12, all students view the Writing prompts on the desktop, laptop, or tablet. The default response mode is keyboarding.

The Writing test is designed to take approximately 45 to 60 minutes. For all grade-level clusters, the Tier B/C Writing tests have recommended timing guidelines for Parts A, B, and C of 10, 20, and 30 minutes, respectively. Note that the approximate test administration time does not include convening students, taking attendance, distributing, and collecting test materials, or explaining test directions, including the directions and practice that precede the test.

3.2.4.1 Writing Test Tiers

Student performance on the Listening and Reading tests determines the appropriate tier that the student will take in the Writing and Speaking tests. For grades 4–12, the test engine automatically routes students to the appropriate tier for Writing. For grades 1–3 Writing, once the students have completed the Listening and Reading tests, test coordinators run a Tier Placement Report that identifies the Writing tier each student is assigned to take. This report is necessary for test administrators to know which tier Writing form to administer to which student, since the Writing test for grades 1–3 is entirely paper based (see Section Writing Tasks for more information about the design of the Writing test). The Writing test has two tiers: A and B/C. In grades 1–3, students must be tested in groups organized by grade-level cluster and tier.

3.2.5 Speaking Test Administration

The Speaking test (including directions and practice) is designed to take approximately 30 minutes. All students in grades complete the ACCESS for ELLs Speaking test on desktop, laptop, or tablet.

Recording response time on every task on the Speaking test has a preset time limit, which varies depending on the grade-level cluster, tier, and task level. Students learn about the time limits in the test directions and practice. Students see a circle change color and then disappear as the time to respond elapses. While there is a limit to how long students can take to record their response, students can navigate the directions, practice, and test items at their own pace. Students click the Next button when they are ready to move on from a screen, without time limits. The test does not advance automatically.

3.2.5.1 Speaking Test Tiers

For each grade-level cluster, the Speaking test has three different tiered forms, Pre-A, A, and B/C. For all grade-level clusters, the tier the student takes is determined by the student's Listening and Reading test results; the test engine automatically routes students to the appropriate Speaking tier. The Pre-A tier is designed to address the needs of newcomer students and to allow those students at the beginning stages of English language development an opportunity to respond to tasks appropriate to what they can do. Tier Pre-A also includes a simplified version of the Speaking test practice to ease the burden of learning how to respond to Speaking tasks on the screen for newcomer students. Most students are placed in either Tier A or Tier B/C.

3.2.5.2 Group vs. Individual Delivery

The Speaking test is administered to small groups of students. For students in all grade-level clusters taking the Tier A and Tier B/C forms, it is recommended that the Speaking test be administered to groups of three to five students.

It is recommended that students taking the Pre-A form be administered the test individually so test administrators can provide additional support during the test. For students in all tiers, the Speaking test may be administered individually or in smaller groups of students as mentioned above, if needed. Test administrators use their professional judgment to consider whether students with high test anxiety or students requiring extra support should be given the test individually or in a very small group.

3.2.6 Test Security

Every effort is made to keep the test secure at all levels of development and administration. WIDA, CAL, and DRC (the entity responsible for printing, distributing, collecting, and scoring the printed tests) follow established policies and procedures regarding the security of the test, and every individual involved in the administration of ACCESS, from the district level to the classroom level, is trained in issues of test security.

All materials for ACCESS for ELLs are considered secure test materials. All users of the WIDA Secure Portal are prompted to read and sign a Nondisclosure and User Agreement upon their first login. Use of the WIDA Assessment Management System and INSIGHT test engine are also subject to the terms of use outlined in the WIDA Assessment Management System. Users are prompted to agree with the test security policy upon their first login. The security of all test materials must be maintained before, during, and after the test administration. Under no circumstances are students permitted to handle secure materials before or after test administration. Test materials should never be left unsecured. The test coordinator should track each secure booklet (i.e., the grades 1–3 Writing test booklets and any grades 4–12 handwritten student response booklets) on the ACCESS for ELLs Security Checklist. Individuals are responsible for the secure documents assigned to them. Secure documents should never be destroyed (e.g., shredded, thrown in the trash) except for soiled documents, which must be destroyed in a secure manner. District and school personnel carrying out their roles in the delivery of this assessment must follow ACCESS for ELLs District and School Test Coordinator Manual guidelines to maintain test security. Test security policies are stated in the Test Policy Handbook for State Education Agencies, available on the WIDA Secure Portal, and the Memorandum of Understanding (MOU) with states.

3.3 Fairness and Accessibility

The WIDA Accessibility and Accommodations Framework provides support for all ELs, as well as targeted accommodations for students with individualized education plans (IEPs) or 504 plans. These supports are intended to increase the accessibility for the assessments for all ELs (see the [ACCESS for ELLs Accessibility and Accommodations Manual](#) for detailed information). Fairness and accessibility are considered throughout the assessment process (i.e., test design, test development, item selection, forms creation, and test administration). For details, please refer to the universal design principles throughout test and item design to the Test and Item Design Plan ACCESS for ELLs Online Annual Summative Assessment and WIDA Screener Online, available in the SEA Secure Portal.

3.3.1 Support Provided to All ELs

Universal design. ACCESS for ELLs incorporates universal design principles to provide greater accessibility for all ELLs. The test items are presented using multiple modalities, including supporting prompts with appropriate animations and graphics, embedded scaffolding, tasks broken into chunks, and modeling that uses task prototypes and guides. These aspects of universal design are built into CAL's item specifications and item review checklists, and CAL test development managers train the CAL language testing specialists on these principles of universal design through training on the use of the specifications and checklists.

Administrative considerations include adaptive and specialized equipment or furniture, alternative microphone, familiar test administrator, frequent or additional supervised breaks, individual or small group setting, monitoring of the placement of responses in the test booklet or on screen, participation in different testing formats (Paper vs Online), reading aloud to self,

specific seating, short segments, verbal praise or tangible reinforcement for on-task or appropriate behavior, and verbal redirection of students' attention to the test (in English or native language).

Universal tools are available to all students taking ACCESS for ELLs to address their accessibility needs. These may either be embedded in the online test or provided by test administrators during testing. The universal tools provided on ACCESS Online are described in Section 3.3.2. The Test Demo videos available on WIDA's [Sample Items page](#) instruct students how to use the universal tools. During testing, students choose whether to use the tools or not, but they are available to all students throughout testing.

3.3.2 Support Provided to ELs with IEPs or 504 Plans

Accommodations include allowable changes to the test presentation, response method, timing, and setting in which assessments are administered. Accommodations are intended to provide testing conditions that do not result in changes in what the test measures; that provide test results comparable to those of students who do not receive accommodations; and that do not affect the validity and reliability of the interpretation of the scores for their intended purposes.

Accommodations are available only to English learners with disabilities when listed in an approved IEP or 504 plan, and only when the student requires the accommodation(s) to participate in ACCESS for ELLs meaningfully and appropriately. Accommodations are delivered locally by a Test Administrator. More information regarding accommodations is provided in the [ACCESS for ELLs Accessibility and Accommodations Manual](#).

WIDA also offers braille and large print accommodations. The braille test is paper based, and the translation and graphics are provided in either contracted or uncontracted braille for Tier B (grades 1–12). This test is used to provide access to the test for ELs who are blind. The Large Print test is used for students with visual impairments. The font size on the large print paper test is increased to 18 point. For the online test, the magnification/zoom tool increases the on-screen font size up to 1.5× or 2×, depending on the size of the computer monitor.

Universal tools are also available to all ELs taking ACCESS for ELLs. Examples of universal tools include highlighter, line guide, magnification, and color overlay. All universal tools are available to all ELs during testing; specific designation is not required prior to testing to make them available to the student during testing. The Test Demo videos available on WIDA's [Sample Items page](#) instruct students how to use the universal tools. During testing, students choose whether to use the tools or not, but they are available to all students throughout testing. Features available during online-based test administration include the following:

- Audio amplification device (provided by student)
- Highlight tool
- Line guide
- Zoom tool (magnifier)
- Sticky notes—which allow students to take notes to prepare responses to Writing items. This tool is only available in the Writing domain.

- Color overlay—which allows students to change the background color that appears behind text, graphics, and response areas. Five colors are available: pink, yellow, blue, green, and orange.
- Color contrast—which allows students to select from a variety of background/text color combinations
- Keyboard shortcuts/equivalents—which are alternatives to using a mouse (for navigating through the test and using online test tools)
- Scratch/blank paper (to be submitted with the test or disposed of according to state policy)

Allowable test administration procedures are variations in standard test administration procedures that provide flexibility to schools and districts in determining the conditions under which ACCESS for ELLs can be administered most effectively. These procedures are available to any student, as needed, at the discretion of the test coordinator (or principal or designee), provided that all security conditions and staffing requirements are met. Examples of allowable test administration procedures include tests administered by familiar school personnel, in an individual or small group setting, in a separate room, with frequent supervised breaks, or in short segments. For detailed information on the allowable test administration procedures, consult the *ACCESS for ELLs Test Administration Manual*.

Schools and districts should consider how accessibility features and allowable test administration procedures can support accessibility to the test for *all* ELs. The accommodations, accessibility features, and allowable test administration procedures are based on (1) accepted practices in English language proficiency assessment; (2) existing accommodation policies of WIDA Consortium member states; (3) consultation with representatives of WIDA member states who are experts in the education and assessment of ELs and students with disabilities; and (4) the expertise of the CAL test developers.

WIDA offers *Alternate ACCESS for ELLs*. This test is intended only for those English learners who have cognitive disabilities that are so significant as to prevent meaningful participation in ACCESS testing, even with accommodations. The results of the Alternate ACCESS for ELLs operational administration appear in a separate technical report.

4. Scoring

4.1 *Multiple Choice Scoring: Listening and Reading*

Listening and Reading items are scored dichotomously, as correct or incorrect. Scale scores for each domain are calculated based on the items administered to the student and the set of those items that the student answers correctly. For details on how scale scores for Listening and Reading are calculated, see Part 2, Chapter 2, "Analysis of Domains."

4.2 *Scoring Performance-Based Tasks: Writing and Speaking*

Trained raters scored student responses to the performance-based tasks in the domains of Writing and Speaking. DRC retains many raters from year to year; the return rater rate was approximately 60% in 2021 and, overall, most raters scoring for ACCESS for ELLs were experienced DRC raters. DRC drew together this pool of experienced raters to staff the scoring pool for ACCESS for ELLs. To complete the rater staffing, DRC accepted application from twenty eight eligible states, all within the Central and Eastern time zones, and then held virtual one on one interviews, during which DRC's recruiting staff screened applications for rater positions. As part of the hiring process, DRC required each candidate to provide an on-demand writing sample, an on-demand math sample, references, and proof of a 4-year college degree. In this screening process, DRC gave preference to candidates who had previous experience scoring students' responses to tasks included in large-scale assessments and candidates with degrees in English language arts. The rater pool consisted of educators, writers, editors, and other professionals with content-specific backgrounds.

Prior to scoring live student responses, the raters underwent thorough training and qualifying. Training was task-specific to ensure that raters understood the nuances of each unique Writing or Speaking task. DRC selected team leaders based on their prior performance as raters and for their leadership skills and assigned them to small groups of raters; typically, there were 7 to 10 raters on each team. The team leaders were responsible for monitoring the performance of their team members and providing ongoing feedback to support accurate scoring. DRC promoted scoring directors, who earned their positions by demonstrating quality work as raters and as team leaders on previous projects, from within. Scoring directors were responsible for a specific set of tasks within a single domain. The scoring directors trained and oversaw the teams of raters assigned to these tasks. What follows are general scoring procedures that DRC utilized.

Preparing Rater Training Materials for Speaking and Writing tasks

CAL test development staff produce materials that DRC uses to train their raters to score ACCESS Speaking and Writing responses. CAL test development staff members who are trained on the Speaking Scoring Scale and the Writing Scoring Scale ("Expert Raters") annually prepare these rater training materials for new Speaking and Writing tasks during field testing.

The Expert Rater begins by reviewing the storyboard for the task (graphics, text, audio script) and by reviewing the anchor responses for an existing task targeting the same grade level cluster, proficiency level, and WIDA ELD standard, in order to internalize the task input and expectations as well as become calibrated to how the Scoring Scale has previously been applied to a similar task. The Expert Rater also reviews documented criteria for anchor responses and score explanations.

Next, the Expert Rater reviews field test responses in DRC ScoreBoard and identifies approximately 5–10 responses per score point. For each response reviewed, the Expert Rater determines the most appropriate score and records any recommendations for potential anchor responses, any questions, or any other observations.

Following the Expert Rater's initial review of responses, the relevant Speaking or Writing Test development manager (TD manager) reviews the responses selected. The TD manager confirms or revises the scores, recording notes and feedback, and finalizes the selection of one anchor response per score point. Anchor responses are typical responses for the grade level cluster and the task, in terms of both the linguistic characteristics and the content of the response. They are clear examples of the score point with both the Expert Rater and the TD Manager agreeing on the score. For the Writing test, for tasks with primarily handwritten responses, the handwriting must also be generally legible to facilitate internalization of the linguistic characteristics by raters.

Once anchor responses are finalized, the Expert Rater writes score explanations for each anchor. Score explanations refer to each dimension of language described in the Scoring Scale descriptors and provide additional explanation with direct quotes from the response to justify why the score point was awarded.

Finally, the TD manager reviews the score explanations to check that they meet the required criteria. The TD manager also selects 20 responses from the initial review to be used as training samples, and reviews and revises any accompanying score notes as necessary. The 20 training samples are selected so that the full range of observed score points are included in the set, and so that the most commonly observed score points for the grade level cluster and tier are well-represented. The TD manager also reviews all notes from the anchor and training sample selection process and, when necessary, compiles any task-specific scoring guidance to be used by raters.

The anchors, explanations, training samples, training sample notes, and any task-specific scoring guidance are then provided to WIDA for review. CAL staff updates the materials as requested by WIDA and delivers the materials to DRC for field test scoring.

Following field test scoring and operational item selection, CAL adds additional training responses to use in rater training for the operational test. The number of training responses for the field test is limited though, with enough responses for the anchor set and one training set, but not enough for operational scoring which requires a second training set and two qualifier sets. This primarily consists of selecting and annotating additional training samples, so that a minimum of 30 samples are provided for operational rater training. In some cases, additional anchor responses are also added to the anchor set, when an appropriate anchor response for

the highest observed score point was not found while preparing for field test scoring but could be identified once a larger pool of scored responses was available.

Rater Training and Qualifying

- DRC assigned each rater a unique ID number and password.
- The scoring director conducted a team leader training session before training the raters. This session followed the same procedures as rater training but was more rigorous and in-depth due to the extra responsibilities required of team leaders. During team leader training, all WIDA materials were reviewed and discussed. To facilitate scoring consistency, it was imperative that all team leaders imparted the same rationale for each response. Once the team leaders were qualified, leadership responsibilities were reviewed, and team assignments were given.
- Rater training began with the scoring director going through the ACCESS for ELLs PowerPoint presentation provided by CAL. The PowerPoint gave scorers a good overview of ACCESS for ELLs and the WIDA scoring process.
- Rater training continued with the scoring director providing an intensive review of the ACCESS for ELLs Scoring Scale, the model student response for Speaking items, and task-specific anchor sets created by CAL. The anchor set contained a collection of student responses that were used to exemplify each possible score point. Each response included a scoring annotation that explained the scoring rationale. Scorers used the ACCESS for ELLs Scoring Scale, the model student response for Speaking, and the anchor sets as primary references during scoring.
- Next, raters practiced by independently scoring responses in training sets. Training sets were created by DRC scoring directors from responses approved by WIDA and CAL. The responses were selected to show raters the range of each score point (e.g., high, mid, and low 2s). This process helped raters recognize the various ways that a student could respond in order to earn each score point outlined and defined in the scoring guidelines. After each training set was taken, the scoring director led a thorough discussion of the responses.
- Once the scoring scale, anchor sets, and training sets were thoroughly discussed, each rater was required to demonstrate understanding of the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the true scores) on at least one of the qualifying sets. Raters who failed to achieve at least 70 percent exact agreement on the first qualifying set were given additional training, either individually or in a small group setting. Raters who did not perform at the required level of agreement by the end of the qualifying process were not allowed to score any student responses. These individuals were removed from the pool of potential raters in DRC's imaging system and released from the project. Qualifying sets were created by DRC scoring directors from responses approved by WIDA and CAL.
- Throughout training, the scoring director provided detailed directions for use of DRC's computerized scoring system and remote communication tools for raters.
- Once raters were trained, qualified, and began live scoring, DRC used recalibration sets and validity responses to keep the raters calibrated on the tasks they were scoring. Recalibration sets were pre-scored sets of responses that were approved by WIDA and

CAL and were used to help refocus raters on WIDA scoring guidelines. Validity responses were also approved by WIDA and CAL and were responses that were pre-scored and used to ensure raters were adhering to WIDA scoring criteria. Recalibration and validity are explained in greater detail below.

Calculating Score Agreement for Score Monitoring

- DRC's handscoring system generated handscoring reports, detailing agreement rates for each rater and task. The reports were automatically generated overnight throughout the course of handscoring and could also be run on demand. DRC provided weekly interrater reliability reports to WIDA throughout the handscoring process to ensure that DRC maintained sufficient quality control throughout the course of scoring.
- For Writing, DRC defines **agreement** as two adjacent scores, reported as %AG (see Section 4.3 for a description of the Writing Scoring Scale). For example, using the Writing Scoring Scale, DRC considers scores of 2 and 2+ as agreement, as well as scores of 2 and 2 or scores of 2+ and 3. However, DRC considers scores of 2 and 3 on the Writing Scoring Scale as **adjacent**, while considering scores of 2 and 3+ as **nonadjacent**.
- For Speaking, DRC defines **agreement** as two scores that are exactly the same, reported as %EX (see Section 4.4 for a description of the Speaking Scoring Scale). Unlike in Writing, where DRC considers two adjacent scores as "Agreement," raters scoring responses to Speaking tasks must demonstrate Exact Agreement (EX) in order to be considered in "agreement."
- WIDA stipulates a minimum interrater agreement rate of 70% for both Writing and Speaking.

Routing Responses to Ensure "Blind" Second Scores

- The DRC scoring system routed and rerouted responses to raters until raters were assigned the prescribed number of scores for all responses. All responses were scored once, and at least twenty percent of the responses were scored a second time. The responses that were used for the twenty percent read- and listen-behinds were randomly chosen by the imaging system at the item level. Additional read- and listen-behinds by the team leaders and scoring directors were done to further ensure reliability. Raters did not see the scores the other raters assigned, and they did not know if they were the first or second rater.
- The purpose of the first and second scores was to monitor interrater reliability by comparing the scores that two separate raters assigned to the same response. When calculating final scores, the first score assigned was the score of record.

Monitoring Scoring (Quality Control)

- Rater accuracy was monitored throughout the scoring session by means of daily and on-demand reports. These reports ensured that an acceptable level of scoring accuracy was maintained throughout the project. Interrater reliability was tracked and monitored with multiple quality control reports. These reports and other quality control documents were generated at the scoring centers, where they were reviewed by the scoring directors, team leaders, and project managers. DRC provided WIDA with access to these reports

on a regular basis throughout the scoring process to provide assurance that the quality control metrics met or exceeded expectations. If a scorer did not meet scoring expectations, a portion of, or all, their scores could be dropped if the scores had not been reported.

- During the handscoring process, the scoring directors communicated regularly with their team leaders to review the statistics generated from the previous day's work, including interrater reliability, score point distributions, and validity reports.
- Throughout handscoring, team leaders conducted routine read- and listen-behinds to observe, in real time, raters' performance. Team leaders utilized live, scored responses to provide ongoing feedback and, if necessary, retraining for raters.
- The DRC system generated interrater reliability reports daily to monitor how often each rater's scores matched other raters' scores, and scoring leaders continually monitored individual rater statistics, comparing them to the group average. If the agreement rate for a rater fell below 70%, supervisors increased monitoring and retraining activities with the rater. If the rater failed to demonstrate improved reliability, DRC released the rater from scoring responses to that task.
- Since the interrater agreement rates were all at or above 70%, the target that WIDA stipulated, the focus turned to raters with lower-than-average agreement rates—even if their agreement rate was at or above 70%. Even when all agreement rates were at or above 70%, scoring supervisors continued to seek opportunities to increase reliability by providing ongoing feedback and retraining raters based on the specific performance of each rater, as evidenced by the quality control reports and observations made when reviewing scores that a rater assigned.
- DRC can retrieve students' responses on demand (e.g., specific grade-level clusters, specific students) should the need arise during or after the scoring process.
- If needed, DRC can re-score a student's response to a task based on task- or response-level information, such as task number, date, score assigned, or rater ID.
- For both Speaking and Writing, DRC used both recalibration sets and validity responses to monitor handscoring quality control. DRC, CAL, and WIDA collaborated to develop these recalibration sets and validity responses. CAL developed an initial pool of responses for use as recalibration and validity checks by selecting responses from a previous administration of the tasks (e.g., a field test). WIDA staff reviewed and approved this pool of responses and their scores. DRC supervisors supplemented this pool of responses as needed by selecting additional responses, which CAL and WIDA approved before use. For each of the first 5 days that raters scored student responses to a task, they scored one recalibration set of five responses. The recalibration sets did not differ from rater to rater. For example, DRC identified a recalibration set to use for the first day that a rater scored students' responses to a specific task; every rater who was working on that task took this same recalibration set on the first day that they worked on that task. After the raters assigned scores to the recalibration set, the scoring director or team leader reviewed the set using descriptors from the scoring scale and the anchor responses to confirm the rationale behind each response's score. Starting on the sixth day that a rater was working on a task, DRC used validity responses to continue monitoring rater performance. DRC seeded the validity responses into the operational

scoring so that the raters did not know which responses were operational and which were validity responses. Reports generated daily compared the scores that each rater assigned to the “true” score for each validity response. When a rater was working on a task, DRC seeded the validity responses in random order into the rater’s queue for scoring. Given enough time, every rater working on a task would score every validity response for that task, but the order in which the raters would see the validity responses would differ.

Handling Unusual Responses

The following processes were in place at DRC to manage specific types of “unusual” responses:

- **Scoring questions.** If a rater had questions about the application of the scoring guidelines to a response (e.g., if they were uncertain as to the proper score that they should assign), the rater forwarded the response to their team leader for assistance. The team leader then reviewed the response with the rater and assigned the proper score. If the rater needed further clarifications, the team leader worked with the rater to review scoring guidelines.
- **Nonscore codes.** Unusual or aberrant responses for which raters could not assign a score based on the scoring guidelines received a nonscorable code (e.g., Writing responses that are entirely blank or consist entirely of scribbles or pictures). DRC’s handscoring team collaborated with WIDA and CAL to define what specifically constituted a nonscorable response to ensure consistency when applying nonscorable codes, and CAL provided this information to DRC along with other task-specific training materials that DRC then used to train its raters. During scoring, when raters assigned a nonscorable code (except for Blank), DRC’s imaging system automatically forwarded the response to a handscoring supervisor for review and approval. If the handscoring supervisor had any questions about the application of non-score codes to specific responses, the supervisor contacted WIDA and CAL representatives for further review and discussion.
- **Alerts.** To handle possible alert responses (i.e., student responses indicating potential issues related to the student’s safety and/or well-being that may require attention at the local level, as well as potential plagiarism and potential teacher interference), DRC’s imaging system gave raters the ability to alert questionable student responses. When a rater flagged a response with the alert status, the imaging system automatically routed the response to handscoring supervisors for review. The states are notified within 24 hours. If the response was related to the student’s safety and/or well-being, and the handscoring supervisors concurred with the alert, it was then forwarded to WIDA’s project management team who provided the response to the appropriate local education agency.
- **Request for originals.** When a rater came across a scanned student response that was difficult to read (for example, having some partially erased text), the rater flagged the response with a “request original” status. If a rater flagged a response as “request original,” DRC’s imaging system automatically forwarded the response to a handscoring supervisor. If the handscoring supervisor agreed that the original student response needed to be reviewed to properly apply the scoring guidelines, the supervisor

forwarded the request to staff in DRC's Operations Services, who located the original student response so the handscoring supervisor could review the response and score it.

Remote Scoring Procedures due to the COVID-19 Pandemic

Prior to 2020, DRC's handscoring centers managed all WIDA handscoring. In 2020, due to the COVID-19 pandemic, DRC shifted from site-based handscoring to remote handscoring to continue meeting all the handscoring deadlines. All WIDA handscoring continued to be remote in 2024. DRC designed the remote scoring to very closely emulate the work carried out in the physical scoring locations. The platform, content, and expectations for quality remained the same. Using a variety of modes of interactive technology (i.e., web screen sharing, webcast, video chat, and chat), DRC conducted rater training and discussions live (virtually). DRC equipped scoring leaders with a variety of tools to ensure that every rater was successful in understanding and applying scoring criteria to student responses.

Remote scoring began with a training session to guide supervisors and raters using the tools that DRC utilized for remote scoring. Once supervisors and raters were trained on the remote scoring process, handscoring commenced for the ACCESS assessments. A description of DRC's remote scoring process follows.

- **System tools—scoring, training, chat.** ScoreBoard is DRC's secure, web-based scoring application that is designed to be used in a distributed environment. The platform is used within DRC's scoring centers and in remote locations (e.g., in a rater's home). Integrated training resources provide the capability to securely maintain digital training materials within the scoring platform itself.
- DRC conducted live, interactive rater training using the Moodle Learning Management System, which mirrored aspects of the scoring room and provided a versatile platform for training. It also served as a place to share files of important documents, including daily scoring statistics and platform user guides. Through embedded communication tools, scoring directors, assistant scoring directors, and team leaders facilitated group and one-on-one training sessions and discussions using audio and video.
- To facilitate instant communication between supervisors and raters, DRC utilized a chat tool called Zulip in conjunction with ScoreBoard and Moodle. Zulip provided a tool for raters to directly ask supervisors questions about responses and allowed supervisors to direct individuals or groups of raters to join Moodle training rooms for important discussions and retraining.
- **Security.** Security is essential to the handscoring process. When users logged into ScoreBoard, they were required to read and accept the security policy before they were allowed to access the project. DRC also required raters to read and sign nondisclosure agreements. During training and large-group discussions, trainers continuously emphasized what security means, the importance of maintaining security, and how all staff accomplish this. In the remote environment, DRC could give these security reminders daily. DRC requires raters working remotely to work in a private environment away from other people (including family members). Raters working in ScoreBoard were not allowed to print from their computers in order to protect the security of the student responses, test questions, and training materials. Restrictions built into ScoreBoard

defined the hours during the day that raters were able to log into the system, ensuring that raters were only scoring responses while supervisors were in place to monitor handscoring and answer any questions.

- **Rater training with Moodle.** DRC conducted rater training remotely as an interactive, comprehensive, hands-on experience. For Writing training, scoring directors trained groups of raters by screensharing PDFs of training materials. Raters individually viewed each training example, with supervisors directing raters to relevant text.
- For Speaking training, scoring directors trained groups of raters by playing the responses aloud over Moodle during live, remote training sessions.
- As with site-based training sessions, supervisors guided the discussion, and raters posed questions to supervisors. The scoring director directed the team leaders and raters to take training and qualifying sets, following the same training flow as they would in the scoring facility.
- **Quality control.** DRC utilized its robust quality control processes and handscoring metrics for all scoring sessions. Scored responses were monitored with second reads, and team leaders conducted read- and listen-behinds. DRC's handscoring system allowed scoring supervisors to determine specific read- and listen-behind rates (frequency of monitoring) for each rater. Any retraining and/or conversations needed because of the monitoring were held in one-on-one video chat sessions. Handscoring quality reports were available daily and on demand for handscoring supervisors and DRC's project leadership, and DRC also provided WIDA staffing with handscoring reports. If a rater fell below 70% exact agreement and failed to improve after retraining and feedback, DRC removed the rater from the project and assigned the responses to other raters to score.

4.3 *Writing Scoring Scale*

The Writing Scoring Scale has six whole score points that range from 1 to 6. The scale descriptors include three different yet interrelated dimensions: discourse, sentence, and word/phrase. These scale descriptors guide raters as they consider all three dimensions to make holistic judgments about which score point best suits a response. The dimensions are distinguished as follows:

- The descriptors for the discourse dimension focus on the degree of organization and the extent to which the response is tailored to the context (e.g., purpose, situation, and audience).
- The descriptors for the sentence dimension evaluate the complexity and grammatical accuracy of sentence structures used in the response.
- The descriptors for the word/phrase dimension specify the range and appropriateness of the original vocabulary used (i.e., text other than that copied and adapted from the stimulus and prompt).

Table 14 shows the Writing Scoring Scale.

Table 14.**WIDA Writing Scoring Scale, Grades 1–12**

| | |
|---|---|
| 5+ | Score Point 6 D: Sophisticated organization of text that clearly demonstrates an overall sense of unity throughout, tailored to context (e.g., purpose, situation, and audience) S: Purposeful use of a variety of sentence structures that are essentially error-free W: Precise use of vocabulary with just the right word in just the right place |
| 4+ | Score Point 5 D: Strong organization of text that supports an overall sense of unity, appropriate to context (e.g., purpose, situation, and audience) S: A variety of sentence structures with very few grammatical errors W: A wide range of vocabulary, used appropriately and with ease |
| 3+ | Score Point 4 D: Organized text that presents a clear progression of ideas, demonstrating an awareness of context (e.g., purpose, situation, and audience) S: Complex and some simple sentence structures, containing occasional grammatical errors that don't generally interfere with comprehensibility W: A variety of vocabulary beyond the stimulus and prompt, generally conveying the intended meaning |
| 2+ | Score Point 3 D: Text that shows developing organization including the use of elaboration and detail, though the progression of ideas may not always be clear S: Simple and some complex sentence structures, whose meaning may be obscured by noticeable grammatical errors W: Some vocabulary beyond the stimulus and prompt, although usage is noticeably awkward at times |
| 1+ | Score Point 2 D: Text that shows emerging organization of ideas but with heavy dependence on the stimulus and prompt and/or resembles a list of simple sentences (which may be linked by simple connectors) S: Simple sentence structures; meaning is frequently obscured by noticeable grammatical errors when attempting beyond simple sentences W: Vocabulary primarily drawn from the stimulus and prompt |
| | Score Point 1 D: Minimal text that represents an idea or ideas S: Primarily words, chunks of language, and short phrases rather than complete sentences W: Distinguishable English words that are often limited to high frequency words or reformulated expressions from the stimulus and prompt |
| D: Discourse Level S: Sentence Level W: Word/Phrase Level | |

When assigning a score, a rater makes an initial judgment about which whole score point (1–6) best describes a response and then determines whether the three descriptors for that whole score point suit that response. If all three descriptors suit the response, the rater assigns the score associated with that score point (e.g., if all three descriptors for score point 3 are appropriate, the rater would assign a score of 3). However, if there is clear evidence that one or two descriptors from an adjacent score point are a better fit, the rater would assign a plus score

between the two applicable whole score points (e.g., if two descriptors for score point 3 seem to fit, but one descriptor for score point 4 is a better fit than the associated descriptor for score point 3, the rater would assign a score of 3+).

In addition to scale descriptors, scoring rules address special cases where responses are nonscorable, completely or partially off task, and completely or partially off topic. These are defined as follows:

- **Nonscorable:** The response is blank; consists only of verbatim copied text; consists only of text that is completely off task; or is entirely in a language other than English; or appears to have been plagiarized from an outside source during testing. More information on how plagiarized responses are handled by DRC is provided in Section 4.2, Handling Unusual Responses.
- **Completely off-task response:** The entire response shows no understanding of or interaction with the prompt. It may be a memorized, previously practiced response or appear to answer another, unrelated prompt. A response that is entirely off task is nonscorable.
- **Completely off-topic response:** The entire response shows a misinterpretation or misunderstanding of the prompt. An off-topic response is related to the prompt but does not seem to address it as intended. However, the response is clearly not a memorized, previously practiced response. Raters score these responses in their entirety using the scoring scale; however, the maximum score for a completely off-topic response is 2+.
- **Partially off-task response:** The response contains both off-task and on-task writing. Raters score these responses by ignoring the off-task portion (which may be memorized and previously practiced) and scoring only the on-task portion using the scoring scale.
- **Partially off-topic response:** The response contains both off-topic and on-topic writing (i.e., a portion of the response shows a misinterpretation or misunderstanding of the prompt). Raters score these responses in their entirety using the scoring scale.

Each student responds to two Writing tasks. One rater assigns a score to each student's response for each task. To calculate a student's total raw score by task, the scores that the raters assigned are converted to whole numbers ranging from 0 to 9, as shown in Table 15. The Writing scoring scale was designed to go up to score point 6. However, we did not have enough responses to estimate the rating scale parameters at each of the 5, 5+, and 6 score points in the empirical data. Therefore, these score points were collapsed into one category for psychometric purposes. Students' scores are then added across tasks, resulting in a total raw score that ranges from 0 to 18.

Table 15.

Rating to Raw Score Conversion (Writing)

| Rating | Raw score |
|-------------|-----------|
| Nonscorable | 0 |
| 1 | 1 |
| 1+ | 2 |
| 2 | 3 |
| 2+ | 4 |
| 3 | 5 |
| 3+ | 6 |
| 4 | 7 |
| 4+ | 8 |
| 5 | 9 |
| 5+ | 9 |
| 6 | 9 |

The ACCESS Writing Scoring Scale is distinct from the WIDA Writing Rubric, which is a tool for evaluating student writing in classrooms and for interpreting student scores from ACCESS Online. CAL and WIDA designed the ACCESS Writing Scoring Scale for trained raters to use to evaluate students' responses to ACCESS writing tasks; thus, it is not appropriate for any other purposes.

4.4 Speaking Scoring Scale

The Speaking Scoring Scale defines five score points: *Exemplary*, *Strong*, *Adequate*, *Attempted*, and *No Response*. The *No Response* score point applies only if the rater uses one of three nonscorable codes: R = dead air or white noise; F = foreign language response; I = nonscorable utterance; K = suspected plagiarism. A nonscorable utterance is defined as one of the following:

- The quality of the audio recording is too poor for any words to be understood. It may be too garbled or too quiet.
- The response contains sounds but no words in English (e.g., *hmmm*, *la la la*, *blah blah blah*).
- The response consists only of a teacher giving instruction or some other overlaying sound (from another student, PA system, etc.).
- The rater believes that the response may have been plagiarized. More information on how plagiarized responses are handled by DRC is provided in Section 4.2, Handling Unusual Responses.

Raters assign scores based on the proficiency level expectations of each task, that is, the level of language proficiency that each task is designed to elicit. The model student response exemplifies these expectations (see Section 2.2.4). In this way, the model response serves as a scoring benchmark. Raters listen to the model response and then score student responses relative to the model. A score of 4 (*Exemplary*) means that the student response demonstrates

English language use that is equal to or beyond the English language use that the model student response illustrates.

Table 16 shows the Speaking Scoring Scale.

Table 16.

WIDA Speaking Scoring Scale

| Score Point | Response Characteristics |
|---|--|
| Exemplary use of oral language to provide an elaborated response | <ul style="list-style-type: none"> • Language use comparable to or going beyond the model in sophistication • Clear, automatic, and fluent delivery • Precise and appropriate word choice |
| Strong use of oral language to provide a detailed response | <ul style="list-style-type: none"> • Language use approaching that of model in sophistication, though not as rich • Clear delivery • Appropriate word choice |
| Adequate use of oral language to provide a satisfactory response | <ul style="list-style-type: none"> • Language use not as sophisticated as that of model • Generally comprehensible use of oral language • Adequate word choice |
| Attempted use of oral language to provide a response in English | <ul style="list-style-type: none"> • Language use does not support an adequate response • Comprehensibility may be compromised • Word choice may not be fully adequate |
| No response (in English) | <ul style="list-style-type: none"> • Does not respond (in English) |

The Speaking Scoring Scale includes descriptors for overall language use, response sophistication, language delivery, and word choice.

Each student responds to three (or six) Speaking tasks, depending upon how the test engine routes the student. A single rater assigns a score to each of those responses, as shown in Table 17. To calculate a total raw score, the scores are then summed, based on the following guidelines:

- For tasks targeting language elicitation at PL 1, there are only three possible score points: *No Response*, *Attempted*, and *Adequate and Above*. This is the case because appropriate responses to PL 1 tasks are single words and short chunks of language, so it is not possible to reliably distinguish between *Adequate*, *Strong*, and *Exemplary* performances.
- For tasks targeting language elicitation at PL 3 and PL 5, each task can be scored on the entire breadth of the scale.
- Each student routed to Tier Pre-A responds to three PL 1 Speaking tasks. Thus, for students in this tier, the total raw score can range from 0 to 6.
- Students routed to Tier A respond to six Speaking tasks, three at PL 1 and three at PL 3. For students in this tier, the total raw score can range from 0 to 18.
- When scoring students' responses to Speaking tasks included in Tier B/C, six points are added to the total raw score, representing a score of *Adequate and Above* for three tasks targeting language at PL 1. Though a Tier B/C student would not be administered any tasks targeting the PL 1 level, it is assumed that a student who had been routed to Tier

B/C would easily achieve a score of *Adequate and Above* on these tasks. Thus, for a student routed to Tier B/C, the total raw score can range from 6 to 30.

Table 17.

Score to Raw Score Conversion (Speaking)

| Score | Raw Score |
|-------------------------------|-----------|
| No Response | 0 |
| Dead air or white noise (R) | 0 |
| Foreign language response (F) | 0 |
| Nonscorable utterance (I) | 0 |
| Suspected plagiarism (K) | 0 |
| Attempted | 1 |
| Adequate/Adequate and Above | 2 |
| Strong | 3 |
| Exemplary | 4 |

DRC trained raters evaluate students' responses to the Speaking tasks using the ACCESS Speaking Scoring Scale. The Speaking Scoring Scale is distinct from the WIDA Speaking Rubric, which is a tool for classroom use and score interpretation. CAL and WIDA designed the ACCESS Speaking Scoring Scale for raters to use to evaluate students' responses to ACCESS speaking tasks; thus, it is not intended to be used for classroom assessment purposes.

5. Summary of Score Reports

5.1 *Individual Student Report*

Score reports (district, school, and student level reports) are made available in WIDA Assessment Management System (AMS) as soon as they are available for each state, and WIDA ships printed reports to school districts and schools at the same time or shortly thereafter. Score reports are available for states to use to identify students' language performance and properly determine language support for ELs. Each state and school district determines when and how students' parents or guardians will receive individual score reports. WIDA provides resources that schools, districts and states may use to aid in score interpretation. (See links below.) How these stakeholders use the material to communicate assessments results is determined locally.

Individual student reports are available in various languages in WIDA AMS, and alternate formats (i.e., Braille or large print) of score reports are available upon request.

WIDA offers several online resources to help communicate test score information to educators, families, and students. (See the [ACCESS for ELLs Score and Reports](#) and the [Family Engagement](#) pages on the WIDA website. A post-testing Q & A webinar about score interpretation is also available on the WIDA Secure Portal.

According to Kim et al. (2016, 2020), educators find interpreting technical information supplied in score reports to be challenging, which suggests a need for more clarity when describing student performance. WIDA plans to convene focus groups to gain an understanding of how various test users (i.e., educators, parents/guardians, students) interpret the information conveyed in current score reports in order to guide efforts to revise those reports for greater clarity.

The Individual Student Report (Figure 20) contains detailed information about the performance of a single student in grades K–12. Its primary users are students, parents/guardians, teachers, and school teams. It provides information about one indicator of a student's English language proficiency: the language needed to access content and succeed in school.

Figure 20.

Individual Student Report



ACCESS for ELLs®
English Language Proficiency Test

Yang, Isabella
Birth Date: | Grade: 04
Tier: A
District ID: WS99999 | State ID: 13118248
School: Training Reports School
District: WIDA Use Only - Sample District
State: WS

Individual Student Report 2025

This report provides information about the student’s scores on the ACCESS for ELLs English language proficiency test. This test is based on the WIDA English Language Development Standards and is used to measure students’ progress in learning English. Scores are reported as Language Proficiency Levels and as Scale Scores.

| Language Domain | Proficiency Level (Possible 1.0-6.0) | Scale Score (Possible 100-600) and Confidence Band See Interpretive Guide for Score Reports for definitions |
|---|---|--|
| | 1 2 3 4 5 6 | 100 200 300 400 500 600 |
| Listening | 2.2 | 283 |
| Speaking | 2.5 | 271 |
| Reading | 2.9 | 334 |
| Writing | 4.7 | 389 |
| Oral Language 50% Listening + 50% Speaking | 2.3 | 277 |
| Literacy 50% Reading + 50% Writing | 4.3 | 362 |
| Comprehension 70% Reading + 30% Listening | 2.7 | 319 |
| Overall* 35% Reading + 35% Writing + 15% Listening + 15% Speaking | 3.6 | 336 |

*Overall score is calculated only when all four domains have been assessed. NA: Not available

| Domain | Proficiency Level | Students at this level generally can... |
|-----------|-------------------|--|
| Listening | 2 | understand oral language related to specific familiar topics in school and can participate in class discussions, for example: <ul style="list-style-type: none">Identify main topics in discussionsCategorize or sequence information presented orally using pictures or objectsFollow short oral directions with the help of picturesSort facts and opinions stated orally |
| Speaking | 2 | communicate ideas and information orally in English using language that contains short sentences and everyday words and phrases, for example: <ul style="list-style-type: none">Share about what, when, or where something happenedCompare objects, people, pictures, and eventsDescribe steps in cycles or processesExpress opinions |
| Reading | 2 | understand written language related to specific familiar topics in school and can participate in class discussions, for example: <ul style="list-style-type: none">Identify main ideas in written informationIdentify main actors and events, in stories and simple texts with pictures or graphsSequence pictures, events or steps in processesDistinguish between claim and evidence statements |
| Writing | 4 | communicate in writing in English using language related to specific topics in school, for example: <ul style="list-style-type: none">Produce papers describing specific ideas or conceptsNarrate stories with details of people, events, and situationsCreate explanatory text that includes details or examplesProvide opinions supported by reasons with details |

SUM-ISR

The Individual Student Report includes four language domain scores (Listening, Speaking, Reading, and Writing) and four language domain composite scores (Oral Language, Literacy, Comprehension, and Overall), as shown in the first table of the score report. In the first column of the last four rows of that table, test users can see how WIDA uses a student's domain scores to calculate each composite score (e.g., for Oral Language, WIDA calculates the composite score based on a student's performance on the Listening and Speaking tests, with scores on each of those tests contributing equally to the composite score). For students who are unable to complete all four domains due to their disabilities, WIDA provides states methods to compute alternative composite scores based on their available domain scores upon request (Sahakyan,2020).

The proficiency level that a student attained in each language domain is presented both graphically and as a whole number followed by a decimal. These are interpretive scores that are based on, but separate from, scale scores. The shaded bar of the graph describes a student's performance in terms of the 6-level English Language Proficiency Scale. The whole number indicates a student's English language proficiency level (1–Entering, 2–Emerging, 3–Developing, 4–Expanding, 5–Bridging, and 6–Reaching) in accordance with the WIDA ELD Standards. English Learners who attain Level 6, Reaching, have moved through the entire second language continuum, as defined by the test and the WIDA ELD Standards.

The decimal indicates the proportion within the proficiency level range that the student's scale score represents, rounded to the nearest tenth. For example, a proficiency level score of 3.5 is halfway between English language proficiency levels 3.0 and 4.0.

To the right of the proficiency level is the reported scale score and the associated confidence band for each domain and composite. A scale score represents a student's performance that has been put on a standardized scale. Students' performance relies on the number of items and item difficulties they respond to correctly. Scale scores allow comparison across different forms and grades. In ACCESS, the scale score ranges between 100–600 in all grades. The confidence band reflects the standard error of measurement for the scale score, a statistical calculation of a student's likelihood of scoring within a particular range of scores if he or she were to take the same test repeatedly without any change in ability. For ACCESS scale scores, the confidence band reflects a 95% probability level.

The second table in the Individual Student Report provides information about the student's proficiency levels expressed as whole numbers. The third column of the table describes what that student should generally be able to do in each of the four language domains, given his or her level of proficiency. For example, as shown in Figure 20, this student received a proficiency level score of 2.5 for Speaking, which suggests that the student should generally be able to "communicate ideas and information orally in English using language that contains short sentences and everyday words and phrases."

If a student was not tested in one (or more) of the language domains, a code of NA (Not Available) will appear in the score report for the impacted language domain(s) and for all composite scores that are calculated using those domain scores. For these students, WIDA provides states with information about statistical methods that can be used to compute alternative composite scores based on a student's available domain scores (Sahakyan, 2020).

When interpreting scores, test users are cautioned to keep in mind these points:

- The report provides information on English proficiency. It does not provide information on a student's academic achievement or knowledge of content areas.
- Students do not typically acquire proficiency in Listening, Speaking, Reading, and Writing at the same pace. Generally,
 - Oral language (L+S) is acquired faster than literacy (R+W).
 - Receptive language (L+R) is acquired faster than productive language (S+W).
 - Writing is usually the last domain to be mastered.
- The students' foundation in their home or primary language is a predictor of their English language development. Those who have strong literacy backgrounds in their native language will most likely acquire literacy in English at a quicker pace than students who do not.
- The Overall score is helpful as a summary of other scores and is used because a single number may be needed for reference. However, it is important to remember that it is compensatory, averaged using weights; a particularly high score in one domain may effectively offset a low score in another domain and vice versa. Similar Overall scores can mask very different performances on individual tests.
- No single scale score or language proficiency level, including the Overall score (composite), should be used as the sole determiner for making decisions regarding a student's English language proficiency. School work and local assessment throughout the school year also provide evidence of a student's English language development.
- Scale scores can be used to make comparisons across grade levels, but not across domains. Each domain has its own score scale, so scale scores should not be used for comparing performance across domains. For example, a scale score of 350 in Listening at grade 3 is not equivalent to a scale score of 350 in Speaking at grade 3. For performance comparisons across domains, proficiency levels should be used.
- Either scale scores or proficiency levels can be used to compare test performance from different years, although it is easier to see changes when examining scale scores.

For detailed information about score reports, please refer to the [ACCESS for ELLs Interpretive Guide for Score Reports](#).

5.2 Other Reports

Student Roster Report. The Student Roster Report contains information on a group of students within a single school and grade. It provides scale scores for individual students in each language domain and composite scores, identical to those appearing in the Individual Student Report. Its intended users are teachers, program coordinators/directors, and administrators.

Frequency Reports. The primary audiences for frequency reports are typically program coordinators/directors, administrators, and boards of education. There are three types of frequency reports:

- School Frequency Report
- District Frequency Report

- State Frequency Report

Each shows the number and percentage of tested students who attained each proficiency level within a given population.

Additional information about the Student Roster Report and Frequency Reports is available in the [ACCESS for ELLs Scores and Reports page](#).