WIDA™
UNIVERSITY OF WISCONSIN–MADISON

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

Series 602, 2023–2024 Administration
Annual Technical Report No. 20B

Executive Summary and Part 1: Purpose, Design, Implementation

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

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 paper-and-pencil version of the ACCESS for ELLs assessment (ACCESS Paper) since the online assessment 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 (ELP) 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 [AERA] 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 reliably and validly assess the English language development of English language learners (ELLs) in grades K–12 according to the WIDA 2012 Amplification of the English Language Development Standards Kindergarten–Grade 12 (WIDA Consortium, 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 Paper Series 602 was administered in school year 2023–2024 in 36 states, the Bureau of Indian Education; the District of Columbia; the Department of Defense Education Activity; the Northern Mariana Islands; and the U.S. Virgin Islands for a total of 41 state entities (henceforth "states").

The Series 602 ACCESS Paper data set used in this report included the results of 555,996 students. The grade with the largest number of students in this report was kindergarten, with 267,037 students, while grade 12 had the fewest, with 8,825 students. Of the participating WIDA states, Florida had the most students (321,509), while the Northern Mariana Islands (MP) had the fewest, with 47 students.

Based on a comparison with prior years' numbers of participating students, 7% more students participated in ACCESS Series 602 testing than the 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 20B) provides technical information pertaining to ACCESS for ELLs Series 602 Paper. A second report (WIDA ACCESS Technical Report 20A) provides technical information for the ACCESS for ELLs Series 602 Online assessment.

Contents

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 of the WIDA Consortium, first in the English Language Proficiency Standards (Gottlieb, 2004; WIDA Consortium, 2007) and then in the amplified 2012 English Language Development (ELD) Standards (WIDA Consortium, 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 English learners

ACCESS for ELLs is offered in two formats: ACCESS Paper, described in this report, and ACCESS Online, 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 the cases of some test items and tasks, the Standards are combined as follows:

- Integrated Social and Instructional Language (SIL), Language of Language Arts (LoLA), and Language of Social Studies (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 for ELLs Paper 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 Consortium, 2012).

The five WIDA ELD Standards should not be thought of in the same sense as content standards (Allen et al., 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.

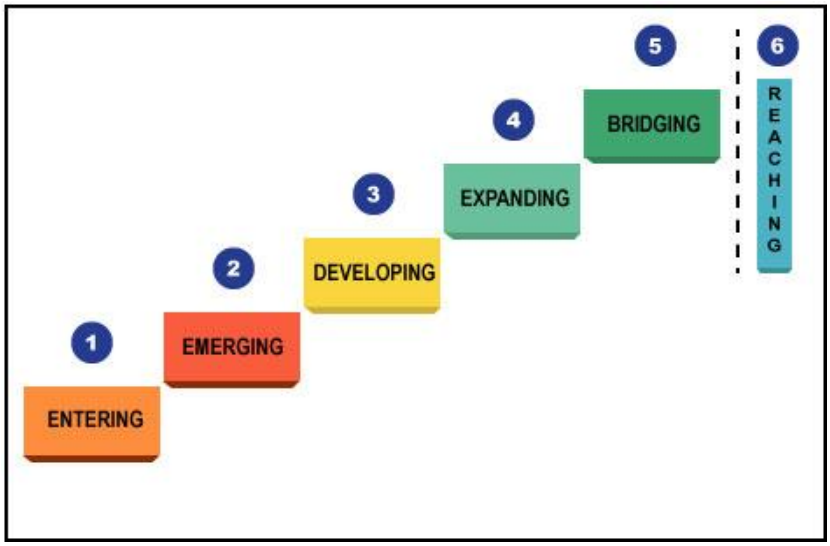
Because of this conceptualization of the WIDA ELD Standards, scores are not reported for each of the Standards, and it is not necessary to assess all five Standards in one domain, as long as each of the Standards is measured on the assessment in some capacity (although ACCESS for ELLs does strive to 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 Consortium, 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 1.3.

Figure 1.3. 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: (a) vocabulary usage at the word/phrase level; (b) language forms and conventions at the sentence level; and (c) 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 refer 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 and 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 ELL students 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 PLs 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 Consortium, 2007) or the amplified 2012 ELD Standards (WIDA Consortium, 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 Consortium, 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 Consortium, 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, 2022). 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 for ELLs Paper is as follows: K, 1, 2, 3, 4–5, 6–8, 9–12.

In the lower grades (grades 1–5), test forms may be shared across grade-level clusters. As described in Section 2.2.1, the Listening and Reading tests were developed prior to the launch of the 2016 operational administration, which represented the shift to the new cluster structure of ACCESS Online. Earlier ACCESS tests had a cluster structure that differed from that of the current ACCESS items in newer developments, in the lower grades. The Speaking and Writing tests were developed using the ACCESS Online cluster structure. ACCESS Paper clusters, therefore, bridge the cluster structure of the older ACCESS assessments and ACCESS Online. For example, the Cluster 2 tests in the domains of Reading and Listening are the same test forms as the Cluster 1 tests. The Cluster 2 tests in the domains of Speaking and Writing are the same test forms as the Cluster 3 tests in these domains. Table 1.5 details the grade-level cluster structure of ACCESS Paper and the shared forms across clusters.

Table 1.5. ACCESS Paper grade-level clusters and shared forms across clusters

ACCESS Paper Grade-level Clusters	Shared Test Forms (Listening and Reading)	Shared Test Forms (Speaking and Writing)	Grade
K	K	K	K
1	Cluster 1 and Cluster 2	Cluster 1	1
2	Cluster 1 and Cluster 2	Cluster 2 and Cluster 3	2
3	Cluster 3 and Cluster 4–5	Cluster 2 and Cluster 3	3
4–5	Cluster 3 and Cluster 4–5	Cluster 4–5	4
4–5	Cluster 3 and Cluster 4–5	Cluster 4–5	5
6–8	Cluster 6–8	Cluster 6–8	6
6–8	Cluster 6–8	Cluster 6–8	7
6–8	Cluster 6–8	Cluster 6–8	8
9–12	Cluster 9–12	Cluster 9–12	9
9–12	Cluster 9–12	Cluster 9–12	10
9–12	Cluster 9–12	Cluster 9–12	11
9–12	Cluster 9–12	Cluster 9–12	12

Note that in our analyses of student participation in the assessment (Part 2, Chapter 1), analysis is conducted by cluster (K, 1, 2, 3, 4–5, 6–8, 9–12). In our analyses of test forms (Part 2, Chapter 2), analysis is conducted at the form level (i.e., in Listening and Reading, a single analysis is conducted for the cluster 1 and cluster 2 form). Test form level analyses are presented for each cluster that the form appears in; if a table of results pertains to more than one cluster, it is repeated in each cluster.

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. Tests must be at the appropriate difficulty level for each 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 that allow 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.

ACCESS Paper test forms are constructed at either Tier A (for students at beginning levels of English proficiency) or Tier B/C (for students at higher proficiency levels). Each grade 1–12 test taker takes either the Tier A form or the Tier B/C form. The Kindergarten assessment is not tiered.

In Listening and Reading, Tier A has items and tasks designed to allow students at the lowest language proficiency levels (PLs 1 and 2) to meet the WIDA ELD Standards at their language proficiency levels, and it includes some items targeted to PL 3. Tier B/C tests include items constructed to target PLs 2 (Emerging) through 5 (Bridging).

In the domain of Writing, Tier A forms include tasks written to elicit language up to PL 3, and Tier B/C forms include tasks written to elicit language up to PL 4 or PL 5. In the domain of Speaking, 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.

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, https://wida.wisc.edu/assess/access/preparing-students/practice.

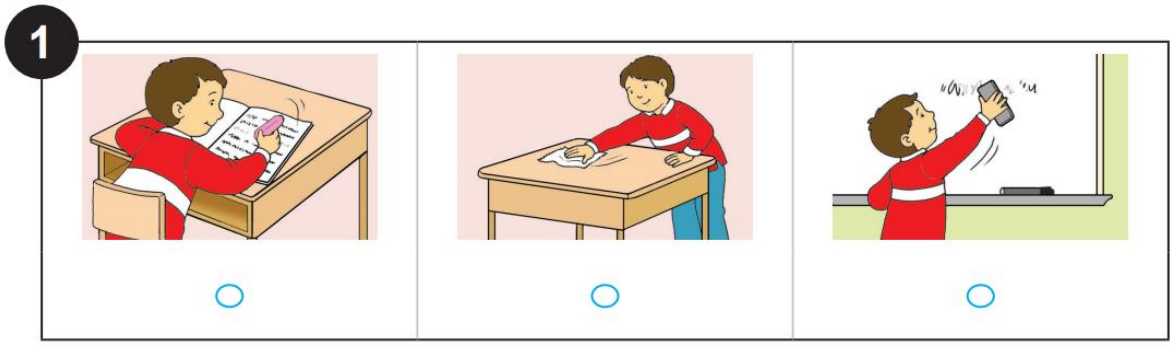
When the task models for ACCESS Paper were first developed, CAL and WIDA addressed issues of fairness by ensuring that principles of Universal Design of Assessments (UDA) (Airhart et al., 2021) were adhered to in this design phase. Therefore, CAL and WIDA collaborated to design the item and task layout on the page to be maximally readable/legible and to contain 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 materials, and to include procedures for accommodation (such as human reader of item stimuli). 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.

Note that this section applies to ACCESS Paper Grades 1–12. For details on the item and task design for Kindergarten, see Section 2.4 and the technical report on the development of the Kindergarten static form (MacGregor et al., 2009).

2.1.1 Listening Items

All Listening items are multiple-choice and are designed to be group administered. They include a prerecorded stimulus passage and question stem. Listening items are multiple-choice items, with one key and two distractors as answer choices. Answer choices are primarily illustrations; 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. Students select their answer by filling in the oval below the response option with a pencil in the test booklet. They can change their answer by erasing the filled-in oval. A sample of Listening item options is provided in Figure 2.1.1.

Figure 2.1.1. Item layout for the ACCESS Paper Listening subtest



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. Folders group together three test items that are written around a common theme, with each item targeting a progressively higher proficiency level.

In ACCESS Paper, the Listening tests have a Tier A and a Tier B/C form for each grade-level cluster; students are placed into the tier based on a decision made at the school or district level as local EL teachers judge students' abilities based on their classroom performance.

Listening items are developed so that each folder appears on a 2-page spread in a test booklet, although some folders go onto a third and fourth page. Scripts containing the item orientation, stimulus, and question stem are audio recorded with professional voice actors and produced by a professional recording studio. Audio playback of test item content is done via audio CD, and explicit instructions on starting and pausing the CD are provided in the Test Administrator Script and the Test Administrator Manual.

Listening items are centrally scored by Data Recognition Corporation (DRC) via an automated process.

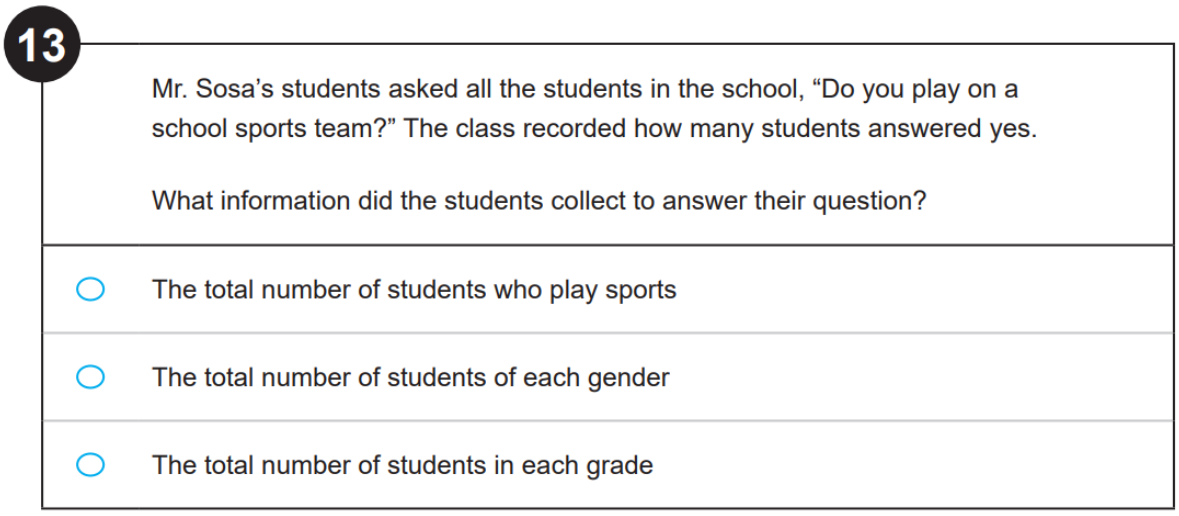
2.1.2 Reading Items

All Reading items are multiple-choice and are designed to be group administered. They are similar in format to Listening items. Reading items are multiple-choice items, with one key and either two or three distractors, depending on grade-level cluster and targeted proficiency level. For grades 1 and 2, all items have a key and two distractors. For grades 3, 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.

The stimulus and question stems for Reading items are written text, and answer choices primarily are also written text, though for grades 1–12 response options for items targeting PLs 1, 2, and 3 may be illustrations rather than text. As with Listening items, Reading items are

grouped into thematic folders of three test items each. In ACCESS Paper, the Reading tests have a Tier A and a Tier B/C form for each grade-level cluster; students are placed into the tier based on a decision made at the school or district level. A sample Reading item is provided in Figure 2.1.2.

Figure 2.1.2. Item layout for the ACCESS Paper reading subtest



Reading items are centrally scored by DRC via an automated process.

2.1.3 Writing Tasks

All Writing tasks are constructed response tasks and are designed to be group administered. Students write responses by hand in paper booklets.

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. As described in Section 2.2.3, DRC raters score students' written responses to these tasks using the entire breadth of the scoring scale; 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 allow students to produce writing samples that fulfill linguistic expectations up to PL 4 or 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 performances may extend above or below the PL 5 range.

In the spirit of providing maximal support and making every provision to ensure that students are allowed to demonstrate the full extent of their written 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 grades 1–5, a word box may be provided, depending on the grade level, targeted proficiency level, and task.

For all grade-level clusters and tiers, the Writing test is group administered by an in-person test administrator. The test administrator reads instructions aloud from the Test Administrator Script and monitors student progress through the test. For all grade-level clusters and tiers, students handwrite their answers in the same test booklet containing the Listening and Reading tests. Figure 2.1.3.1 provides an example of the task layout for the Writing test. Figure 2.1.3.2 provides an example of the accompanying script.

Figure 2.1.3.1. Example task in test booklet for the ACCESS Paper Writing subtest

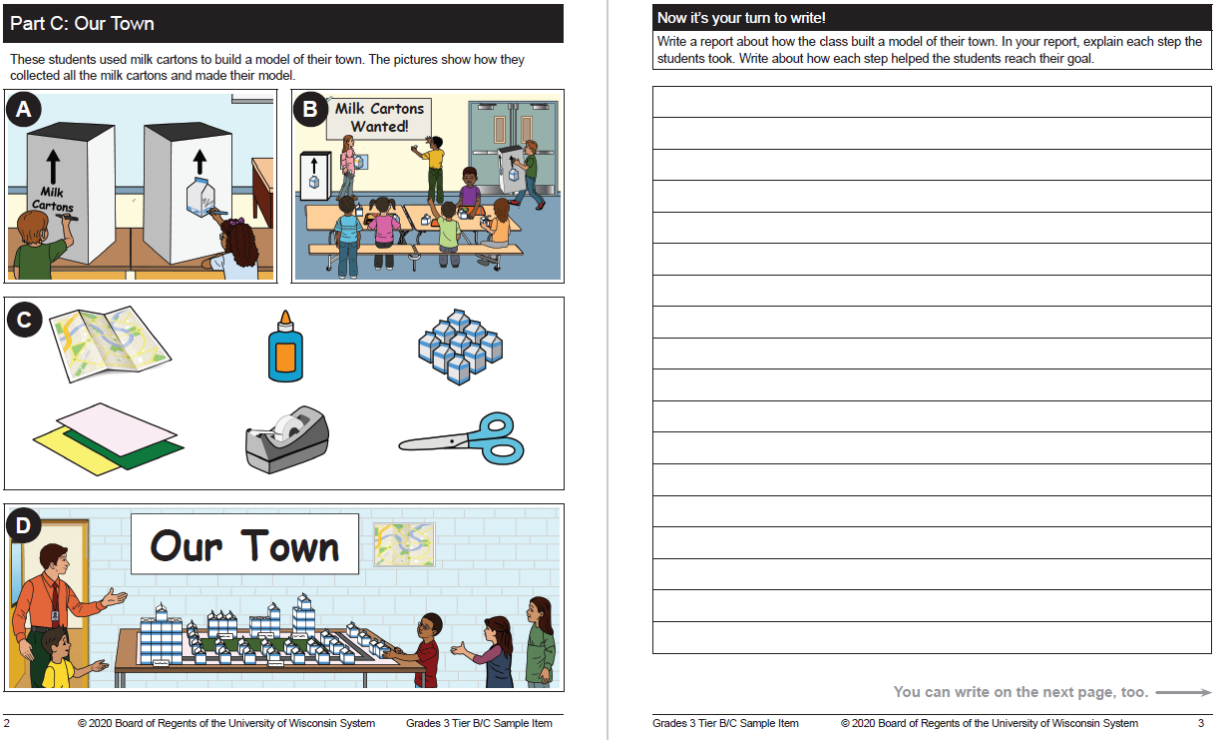
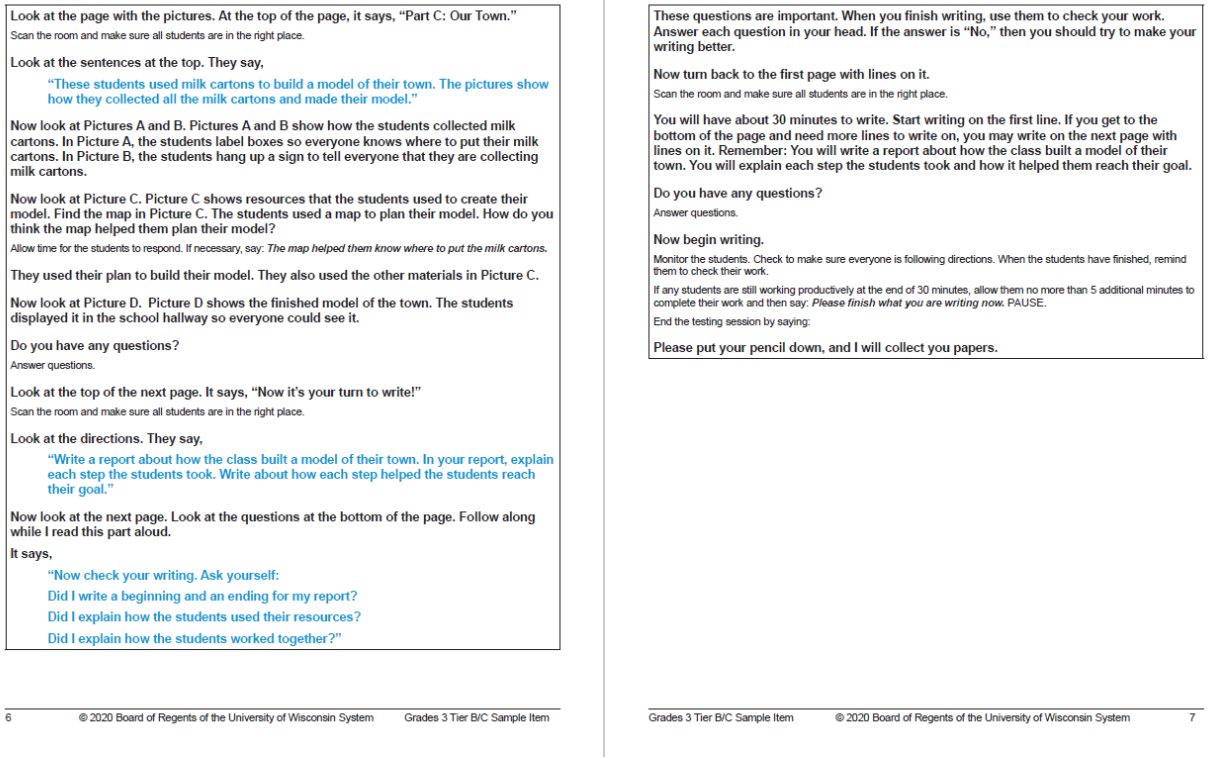


Figure 2.1.3.2. Example script for the ACCESS Paper Writing subtest



Look at the page with the pictures. At the top of the page, it says, "Part C: Our Town."
Scan the room and make sure all students are in the right place.

Look at the sentences at the top. They say,

"These students used milk cartons to build a model of their town. The pictures show how they collected all the milk cartons and made their model."

Now look at Pictures A and B. Pictures A and B show how the students collected milk cartons. In Picture A, the students label boxes so everyone knows where to put their milk cartons. In Picture B, the students hang up a sign to tell everyone that they are collecting milk cartons.

Now look at Picture C. Picture C shows resources that the students used to create their model. Find the map in Picture C. The students used a map to plan their model. How do you think the map helped them plan their model?
Allow time for the students to respond. If necessary, say: *The map helped them know where to put the milk cartons.*

They used their plan to build their model. They also used the other materials in Picture C.

Now look at Picture D. Picture D shows the finished model of the town. The students displayed it in the school hallway so everyone could see it.

Do you have any questions?
Answer questions.

Look at the top of the next page. It says, "Now it's your turn to write!"
Scan the room and make sure all students are in the right place.

Look at the directions. They say,

"Write a report about how the class built a model of their town. In your report, explain each step the students took. Write about how each step helped the students reach their goal."

Now look at the next page. Look at the questions at the bottom of the page. Follow along while I read this part aloud.
It says,

"Now check your writing. Ask yourself:
Did I write a beginning and an ending for my report?
Did I explain how the students used their resources?
Did I explain how the students worked together?"

6 © 2020 Board of Regents of the University of Wisconsin System Grades 3 Tier B/C Sample Item

These questions are important. When you finish writing, use them to check your work. Answer each question in your head. If the answer is "No," then you should try to make your writing better.

Now turn back to the first page with lines on it.
Scan the room and make sure all students are in the right place.

You will have about 30 minutes to write. Start writing on the first line. If you get to the bottom of the page and need more lines to write on, you may write on the next page with lines on it. Remember: You will write a report about how the class built a model of their town. You will explain each step the students took and how it helped them reach their goal.

Do you have any questions?
Answer questions.

Now begin writing.
Monitor the students. Check to make sure everyone is following directions. When the students have finished, remind them to check their work.
If any students are still working productively at the end of 30 minutes, allow them no more than 5 additional minutes to complete their work and then say: ***Please finish what you are writing now.*** PAUSE.
End the testing session by saying:

Please put your pencil down, and I will collect you papers.

Grades 3 Tier B/C Sample Item © 2020 Board of Regents of the University of Wisconsin System 7

2.1.4 Speaking Tasks

The Speaking test is administered individually to each test taker. The test is media delivered. Students listen to an audio recording of the test input while following along in a test booklet.

Stimuli on the Speaking test include graphics, audio, and text, presented in a test booklet as a series of "speech bubbles" from the perspective of the Virtual Test Administrator (VTA) and virtual model student. All text is multimodal, presented both in the test booklet and read aloud on the audio CD. Scripts containing the task content are audio recorded with professional voice actors and produced by a professional recording studio. Audio playback of test item content is done via audio CD, and explicit instructions on starting and pausing the CD are provided in the Test Administrator Script and the Test Administrator Manual.

The CD audio stimuli are presented in terms of a 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. Before responding to each task, students first listen to the model student respond to a parallel task. The purpose of the model is to demonstrate task expectations to both students and to the Test Administrator, who scores the Speaking test. Students respond orally to the tasks, with their responses scored immediately by the Test Administrator using a scoring scale. The Test

Administrator records scores on the Speaking test in the same booklet the student used for the Listening, Reading, and Writing tests.

2.2 Test Design

This section describes how ACCESS Paper is assembled to ensure that the evidence collected is (a) sufficient to make the required decisions based on the test results, and (b) appropriate for the student's level of proficiency. This section provides information on the test design for the two forms of ACCESS Paper (Tier A and Tier B/C) and the design of each form. Note that this section applies to ACCESS Paper Grades 1–12. For details on Kindergarten, see Section 2.4 and the technical report on the development of the Kindergarten static form (MacGregor et al., 2009).

2.2.1 Listening

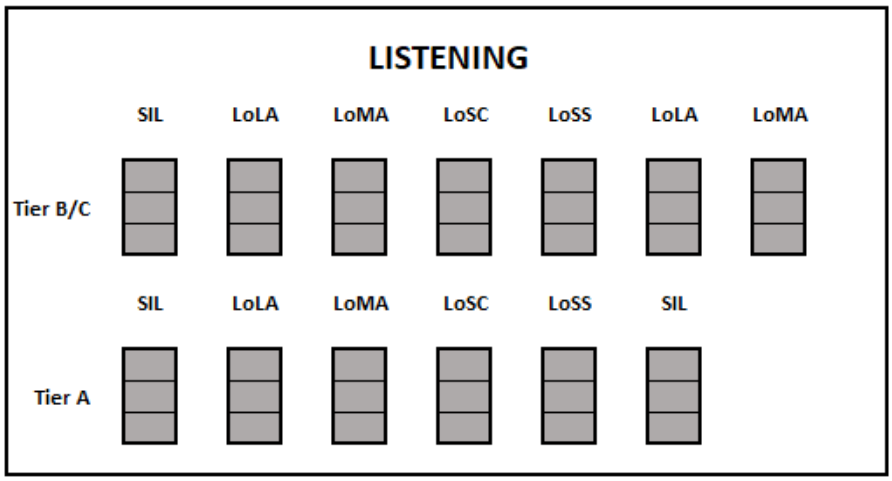
For the ACCESS Listening test, Table 2.1.1 shows, for each test form, the number of items, the targeted range of WIDA proficiency levels, the item types, the response format, and the scoring procedure.

Table 2.2.1. Number and types of items on the Listening test

Grade-Level Cluster	Tier	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
1	A	18	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
1	B/C	21	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
2	A	18	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
2	B/C	21	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
3	A	18	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
3	B/C	21	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
4–5	A	18	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
4–5	B/C	21	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
6–8	A	18	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
6–8	B/C	21	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
9–12	A	18	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
9–12	B/C	21	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored

Figure 2.2.1 presents the Listening test design, showing the distribution of folders by Standard for each tier. In this figure, each small gray box represents an item.

Figure 2.2.1. Distribution of items by Standard for each tier of the Listening test



Note that the test design is slightly different between Tier A and Tier B/C. Tier B/C students, who potentially may be reclassified by the assessment, take a slightly longer test and take two folders each assessing the Language of Language Arts and the Language of Mathematics Standards. Tier A students receive a second folder assessing the Social and Instructional Language Standard, under the assumption that less proficient students will find this Standard more accessible.

Although timing guidance is provided to Test Administrators in the Test Administrator Manual, the Listening subtest is untimed.

2.2.2 Reading

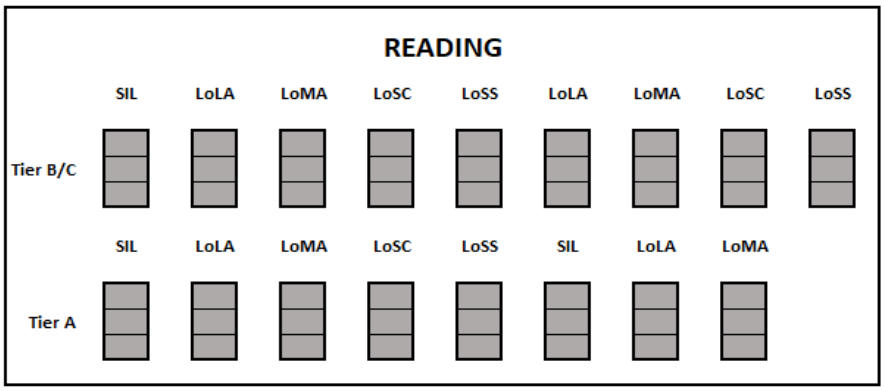
For the ACCESS Reading test, Table 2.2.2 shows, for each test form, the number of items, the targeted range of WIDA proficiency levels, the item types, the response format, and the scoring procedure.

Table 2.2.2. Number and types of items on the Reading test

Grade-Level Cluster	Tier	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
1	A	24	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
1	B/C	27	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
2	A	24	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
2	B/C	27	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
3	A	24	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
3	B/C	27	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
4–5	A	24	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
4–5	B/C	27	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
6–8	A	24	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
6–8	B/C	27	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored
9–12	A	24	PL1–PL4	Multiple choice	Dichotomous selected response	Machine scored
9–12	B/C	27	PL2–PL5	Multiple choice	Dichotomous selected response	Machine scored

Figure 2.2.2 presents the Reading test design, showing the distribution of folders by Standard for each tier. In this figure, each small gray box represents an item.

Figure 2.2.2. Distribution of items by Standard for each tier of the Reading test



As with Listening, the Reading Tier A test is shorter and focuses on Standards deemed more accessible for lower-proficiency students.

Although timing guidance is provided to Test Administrators in the Test Administrator Manual, the Reading subtest is untimed.

2.2.3 Writing

For the ACCESS Writing test, Table 2.2.3 shows, for each test form, the number of tasks, the targeted range of WIDA proficiency levels, the task types, the response format, and the scoring procedure.

Table 2.2.3. Number and types of items on the Writing test

Grade-Level Cluster	Tier	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
1	A	4	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
1	B/C	3	PL2–PL5	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
2	A	3	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
2	B/C	3	PL2–PL5	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
3	A	3	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
3	B/C	3	PL2–PL5	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
4–5	A	3	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
4–5	B/C	3	PL2–PL5	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC

Grade-Level Cluster	Tier	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
6–8	A	3	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
6–8	B/C	3	PL2–PL5	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
9–12	A	3	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
9–12	B/C	3	PL2–PL5	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC

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, some students will score throughout the proficiency range.

Except for grade 1 Tier A, both tiers consist of three tasks. Grade 1 Tier A has four tasks, designed specifically to allow beginning writers at this grade to demonstrate their ability in the domain of Writing. Figure 2.2.3.1 and Figure 2.2.3.2 present the Writing test design, showing the distribution of tasks for each tier. In these figures, each colored box represents a task. The number in the box represents the targeted proficiency level of the task.

Although timing guidance is provided to Test Administrators in the Test Administrator Manual, the Writing subtest is untimed.

Figure 2.2.3.1. Distribution of tasks by targeted proficiency level for each tier of the grade 1 Writing test

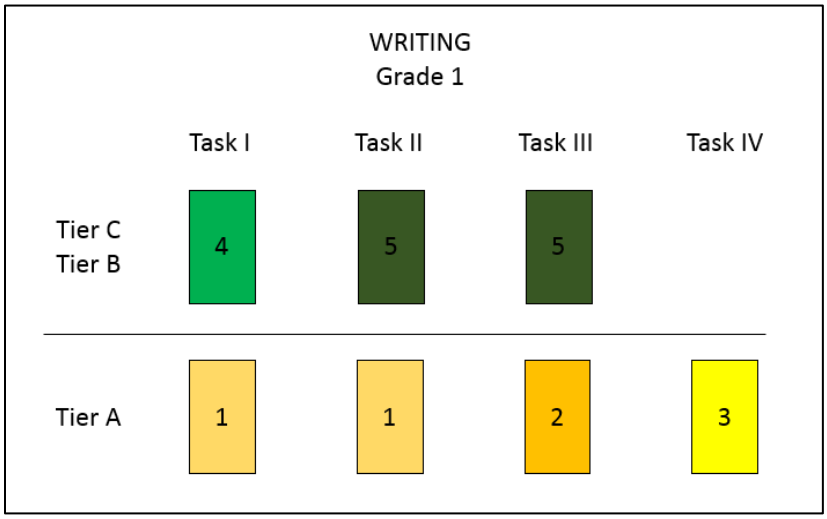
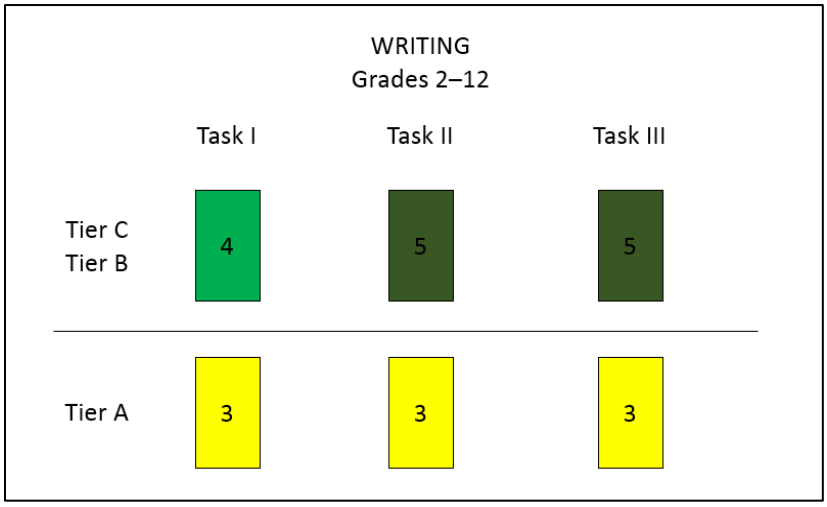


Figure 2.2.3.2. Distribution of tasks by targeted proficiency level for each tier of the grades 2–12 Writing test



2.2.4 Speaking

For the ACCESS Speaking test, Table 2.2.4 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.

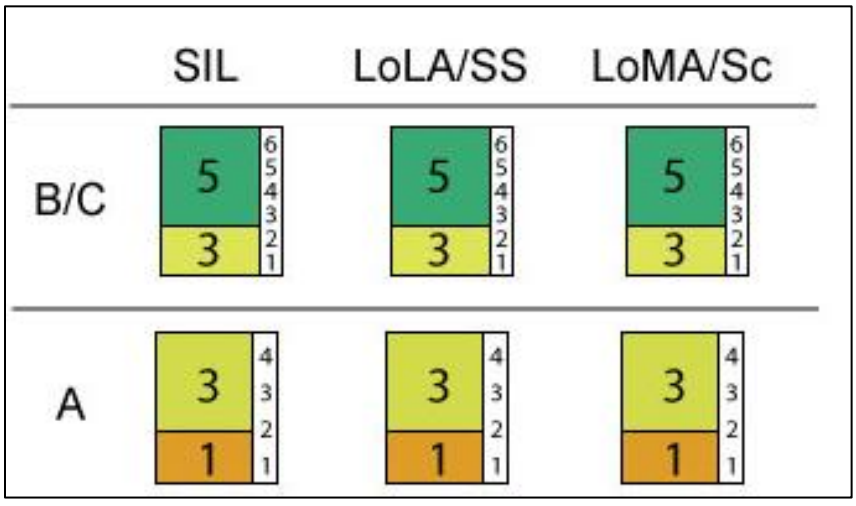
Table 2.2.4. Number and types of items on the Speaking test

Grade-Level Cluster	Tier	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
1	A	6	PL1–PL3	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
1	B/C	6	PL3–PL5	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
2	A	6	PL1–PL3	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
2	B/C	6	PL3–PL5	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
3	A	6	PL1–PL3	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
3	B/C	6	PL3–PL5	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
4–5	A	6	PL1–PL3	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
4–5	B/C	6	PL3–PL5	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator

Grade-Level Cluster	Tier	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
6–8	A	6	PL1–PL3	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
6–8	B/C	6	PL3–PL5	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
9–12	A	6	PL1–PL3	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator
9–12	B/C	6	PL3–PL5	Speaking constructed response	Polytomous constructed response	Human scored; scored by Test Administrator

Figure 2.2.4 shows the format of the Speaking test. The Speaking test includes tasks that target language elicitation at three proficiency levels: 1, 3, and 5. The tasks are grouped into thematic folders, which are aligned to one or two of the WIDA Standards. These folders are generally presented in the same order as the folders in the Listening and Reading tests; folders aligned to SIL are presented first, then folders aligned to LoLA, and then folders aligned to LoMa.

Figure 2.2.4. Distribution of tasks for each tier of the Speaking test



As shown in Figure 2.2.4, the Speaking test includes two tiers. 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. For example, the Tier A and Tier B/C folders that address Social and Instructional Language in each grade cluster make up a single thematic panel, with the PL 3 tasks shared across tiered folders in a panel. In other words, within a Social and Instructional Language panel, the same PL 3 task appears on both the Tier A and the Tier B/C form.

Although timing guidance is provided to test administrators in the Test Administrator Manual, the Speaking subtest is untimed.

2.3 Test Construction

2.3.1 Item and Task Development

ACCESS Paper Series 503 is one of two static rotating Paper test forms. The ACCESS testing program transitioned in 2016 from an entirely paper-based program to the launch of ACCESS in both Online and Paper formats.

The CAL TD team developed the Listening and Reading items for ACCESS Paper when ACCESS was entirely paper based, prior to the launch of ACCESS Online. The CAL TD team also developed most ACCESS Paper Writing tasks for ACCESS when it was entirely paper based; however, a small subset of Writing tasks on ACCESS Paper Series 503 were developed as online tasks that were subsequently reformatted for administration as paper-based tasks. The CAL TD team developed the Speaking tasks and field tested them as ACCESS Online tasks before being reformatted for administration as ACCESS Paper tasks.

The general process of item writing and editing, and of item Bias & Sensitivity and Content reviews, remained similar across these transitions. For ACCESS Paper items and tasks, trained item writers worked from item/task specifications to draft items and tasks within the thematic folder design. After item writing was complete, the CAL TD team reviewed the folders, using a standard checklist, to determine which would undergo further development and which would be retired. Folders then went to their first external review, the Standards Expert review.

During the Standards Expert review, educators provided feedback about the overall grade-level appropriateness of the language and content of the items and tasks to ensure that no drift, in terms of grade-level appropriateness of the content or the language, occurred between the content generated during item writing and what was intended in the specifications. CAL recruited educators with ESL and content-area expertise to serve as Standards Experts and provided synchronous training on how to conduct the review and complete the review questionnaire. CAL Language Testing Specialists prepared a short questionnaire with open-ended questions about each folder and sent the questionnaires and folders to the Standards Experts.

Subsequent to the Standards Expert review, all content proceeded through a rigorous folder refinement stage internal to CAL. Folder refinement included numerous steps, including additional research and sourcing/fact-checking, meticulous review against a comprehensive, industry-standard item development checklist with peer review that other Language Testing Specialists carried out, as well as review by the Test Development Manager and the Director of Test Development and successive rounds of revision before sign-off. During this stage, all aspects of the items and tasks were scrutinized: the WIDA proficiency level of the stimulus, the graphic support, the question stems, and response options (for the Listening and Reading tests), and the task prompts (for the Speaking and Writing tests). The CAL TD team also conducted mock administrations. During this phase, Language Testing Specialists produced other ancillary materials, such as Test Administrator scripts. Upon sign-off, the CAL TD team worked with the CAL Production team to generate the graphics used on the test. Once the graphics had been generated, they were inserted into the folders, and layout review and fact-

checking were conducted (with Test Development Manager sign-off) to ensure that the items and tasks were ready for external Content Review and Bias & Sensitivity Review.

Content Review and Bias & Sensitivity Review are external reviews that educators and WIDA staff carry out on ACCESS items and tasks. WIDA assembles 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 & Sensitivity Reviews. Educators conduct Content Reviews by grade-level cluster (G1, G2-3, G4-5, G6-8, and G9-12). The educators who are recruited to review a particular grade-level cluster's content (4 reviewers per cluster) have experience teaching English language learners and are either currently teaching that grade-level cluster or have extensive prior experience teaching students who are in that 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 a) consortium members, b) school locale (rural/suburban/urban), and c) 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 & 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 & 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 group'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 grouping or have extensive prior experience teaching such students. 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 a) consortium members, b) school locale (rural/suburban/urban), and c) years of teaching experience. The CAL TD team and WIDA TD staff conduct training for all new and returning reviewers before any items and 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 conducts a separate, asynchronous review around the time of the Content Review and Bias & Sensitivity Review, using the same materials that the educators review, and provide written feedback on the materials.

The CAL Language Testing Specialists compile all Content Review and Bias & 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 team then revises the graphics. The input and feedback from educators at various stages in the item/task development process served as evidence that each item or task was appropriate for the age and grade-level cluster for which it was intended.

Tasks in the domain of Writing and Speaking underwent one additional step: two rounds of small-scale tryouts with educators and students. These tryouts allowed the CAL TD team to evaluate whether each Speaking and Writing task would effectively elicit language at its targeted WIDA proficiency levels. In the initial round of tryouts, members of the CAL TD team recruited schools to permit CAL staff to administer the tasks to, and conduct cognitive labs with, students with consent to participate. The tasks were then revised and subjected to a second round of tryouts, this time conducted by classroom teachers with their students, who were also recruited by CAL and WIDA to participate. CAL Language Testing Specialists used the results, including student responses, cognitive lab observations of students, and student and teacher feedback, to inform final revisions to the tasks prior to field testing.

After the CAL Language Testing Specialists completed edits from the Content Review and Bias & Sensitivity Review (and tryout edits for Speaking and Writing), they then prepared the folders for final production. Additionally, they produced audio recording scripts for professional audio recording, arranged for recording the audio files, completed extensive quality control checks for both content and technical specifications of the audio (e.g., file types, recording quality, and compression levels), conducted final layout reviews, and performed key checks for the Listening and Reading tests. WIDA signed off on all materials prior to administration. Items and tasks that reached this point then went through field testing and test assembly processes, described in the next subsection by domain.

Throughout item development, the CAL TD team focused on issues of fairness. First, the team applied principles of Universal Design of Assessments (UDA) during item development. At the item/task specification level, the CAL TD team aimed to precisely define the construct that each item or task was intended to measure. For the linguistic content of items and tasks, several principles for UDA were built into the item development checklists and were specifically reviewed by CAL's TD managers and external reviewers (including WIDA staff and outside educators during Standards Expert review and Bias & Sensitivity and Content reviews), including:

- Inclusive assessment population
- Precisely defined constructs
- Accessible, nonbiased items and 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. Through maintaining a focus on fairness throughout the test development cycle by integrating the principles of UDA in various steps, the CAL TD team ensured that ACCESS Paper items and tasks were best positioned to be maximally fair for all populations.

Note that this section applies to ACCESS Paper Grades 1–12. For details on Kindergarten, see Section 2.4 and the technical report on the development of the Kindergarten static form (MacGregor et al., 2009).

2.3.2 Field Testing and Item Selection

2.3.2.1 Listening and Reading

The Listening and Reading items for ACCESS Paper were created by the CAL TD team before the launch of ACCESS Online when ACCESS was entirely paper-based. ACCESS was first field tested in 2004, and from 2004 to 2014, development continued for ACCESS, culminating in Series 303, operational in 2014–2015. For further detail on this original field test and on the processes for ongoing item development from 2004 to 2014, see Section 2.3.1, along with the ACCESS for ELLs Technical Reports, particularly ACCESS for ELLs Technical Report No. 1, *Development and Field Test of ACCESS for ELLs* (Kenyon, 2006), and *Annual Technical Report for ACCESS for ELLs® English Language Proficiency Test, Series 303* (CAL, 2016b).

In all grade-level clusters, the Tier A Listening and Reading forms are static forms, which were constructed prior to the launch of ACCESS Online.

In all grade-level clusters, the operational Tier B/C forms in Listening and Reading forms for Series 602 are identical to those administered in Series 501. These forms are composed of items that were previously operational in Series 400 and 401 and that were developed, as described in Section 2.3.1 above, during the development cycles when ACCESS was entirely paper based. Beginning with Series 403, to streamline operational administration, CAL and WIDA decided to combine ACCESS Paper Listening and Reading Tier B and Tier C tests to create a new Tier B/C test in Listening and in Reading for each grade-level cluster.

2.3.2.2 Writing

There are two static rotating forms for ACCESS Paper Writing. The first of these is composed of the same set of items, across all grade-level clusters and tiers, as the test used the first year of ACCESS Online. The second form is composed of the same set of items, across all grade-level clusters and tiers, as the test used the second year of ACCESS Online.

Tasks on the first of the two rotating static forms were used operationally prior to the launch of ACCESS Online and were re–field tested in the Online mode for the first year of ACCESS Online. Tasks selected for use in the first ACCESS Online operational test were then reformatted for presentation in the first of the Paper static forms.

The second rotating static form uses continuing tasks from the first form, as well as tasks newly field tested for the second year of ACCESS Online and then reformatted for Paper presentation. For further details on this field test, see the *Series 401 Online ACCESS Technical Report* (CAL, 2018).

ACCESS Paper 602 is the second of the two rotating static forms.

2.3.2.3 Speaking

The Speaking test for ACCESS Paper is likewise one of two static rotating forms. The first of these forms is composed of the same set of items, across all grade-level clusters and tiers, as the second year of the ACCESS Online Speaking test. The second form is composed of the same set of items, across all grade-level clusters and tiers, as the third year of the ACCESS Online Speaking test. Speaking tasks have some differences in presentation between Online and Paper. In addition, the Paper test does not include the Speaking tier Pre-A, which is included in the Online test.[1]

Tasks for these two rotating forms were field tested during the initial ACCESS Online field test, as well as through embedded field testing during the first and second years of the ACCESS Online assessments. These Speaking tasks went through both quantitative and qualitative analyses following the field test to determine their appropriateness for inclusion in the next year's operational test. After field testing, the Speaking tasks were then produced in the paper-based format.

[1] Students with very low ability levels in the Listening and Reading domains are routed to the Pre-A tier in ACCESS Online Speaking. The purpose of the Pre-A tier is to reduce the affective impact of the test on these students. As the Paper test is not adaptive, there is no way to route these students to Pre-A for Paper.

2.4 Kindergarten

Kindergarten ACCESS for ELLs is a static form and is not refreshed from year to year.

2.4.1 Test Design

CAL and WIDA designed Kindergarten ACCESS for ELLs to be engaging for very young children, and the test design was informed by consultation with kindergarten teachers and a panel of early childhood assessment experts. The test design incorporates a high-interest, age-appropriate storybook format, using child-friendly graphics, and includes manipulatives for students to demonstrate comprehension. The test is built on two thematic texts in a storybook format, one narrative and one expository. The Test Administrator reads the story aloud. There are Listening, Speaking, Reading, and Writing assessment tasks related to each text. To minimize testing times and to ensure that students are presented with assessment tasks appropriate to their abilities, the test includes stopping rules (designed to ensure that children of beginning proficiency are not overchallenged) and skipping rules (designed so that children of more advanced proficiency can skip forward to more challenging tasks).

The test is administered one-on-one by trained Test Administrators, who mark responses in the Student Response Booklet.

Table 2.4.1 provides, for each domain, the number of items, the targeted range of WIDA proficiency levels, the item types, the response format, and the scoring procedure.

Table 2.4.1. Number and types of items on Kindergarten ACCESS

Domain	Number of Items	Targeted PL Range	Item Types	Response Formats	Scoring Procedures
Listening	30	P1–P5	Dichotomous	Student points to picture or manipulates cards	Administrator records response (correct/incorrect) in Student Response Booklet
Speaking	10	P1–P5	Dichotomous	Oral response	Administrator records response (correct/incorrect) in Student Response Booklet
Writing	6	P1–P5	Dichotomous and Polytomous	Student handwrites in booklet	Administrator records response (correct/incorrect) for dichotomous tasks and rates responses and records rating for polytomous tasks
Reading	30	P1–P5	Dichotomous	Student reads aloud or matches picture cards with text cards	Administrator records response (correct/incorrect) in Student Response Booklet

2.4.2 Test Construction

Field testing for Kindergarten ACCESS was conducted in 2008. A full description of item development, field testing, final forms selection, and initial standard setting for Kindergarten ACCESS can be found in the technical brief *Development and Field Test of Kindergarten ACCESS for ELLs* (MacGregor et al., 2009). Cut scores for Kindergarten were most recently updated in the 2016 ACCESS standard setting (Cook & MacGregor, 2017); see Part 2, Section 2.1 for more information.

2.4.3 Item and Task Design

As noted previously, the Kindergarten ACCESS test is composed of two thematic texts, one narrative and one expository. The items and tasks are designed to build upon the content of these texts.

In the domain of Listening, the Test Administrator reads the prompt aloud to the student, and the student responds by either pointing to an item in a picture or manipulating a picture card.

The Test Administrator records the response (correct or incorrect) in the Student Response Booklet.

Students respond to Writing tasks in the Student Response Booklet. The initial Writing tasks for each thematic text are dichotomously scored by the Test Administrator. The Test Administrator Script indicates the level required for a task to meet expectations and to be scored as "correct". The Test Administrator scores the final Writing task in each thematic text section using a rating scale. The Test Administrator rates the student's Writing on a scale of 0 to 6.

The Test Administrator reads the Speaking tasks aloud, and students respond orally. The Test Administrator dichotomously scores the tasks. The Test Administrator Script indicates the level required for a task to meet expectations and to be scored as "correct".

To administer Reading tasks, Test Administrators ask students to identify letters or read text. Students respond by manipulating picture cards or by pointing at pictures. Students may also read aloud. The Test Administrator records the response (correct or incorrect) in the Student Response Booklet.

The items on Kindergarten ACCESS were developed to collectively assess all five WIDA Standards in all domains across the proficiency levels. To keep the test an appropriate length for the population, it was not possible to assess each Standard at each proficiency level in each domain. Therefore, tasks were distributed by Standard across the proficiency levels and domains to achieve appropriate coverage. Table 2.4.3.1 through Table 2.4.3.8 show the number of items per domain, proficiency level, and WIDA Standard for each of the thematic texts, narrative and expository.

Although the average time per test is provided to Test Administrators in the Test Administrator Manual, Kindergarten ACCESS is untimed.

Student Response Booklets are centrally scanned at DRC.

Table 2.4.3.1. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Listening: Narrative storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4	Number of items at targeted PL 5
SI	3	0	3	0	3
LA	0	0	0	0	0
MA	0	0	0	0	0
SC	0	0	0	0	0
SS	0	3	0	3	0

Table 2.4.3.2. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Listening: Expository storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4	Number of items at targeted PL 5
SI	3	0	0	0	0
LA	0	0	0	3	0
MA	0	3	0	0	0
SC	0	0	0	0	0
SS	0	0	3	0	3

Table 2.4.3.3. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Speaking: Narrative storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4	Number of items at targeted PL 5
SI	0	0	3	0	0
LA	0	0	0	0	3
MA	0	0	0	0	0
SC	0	0	0	0	0
SS	3	3	0	3	0

Table 2.4.3.4. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Speaking: Expository storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4	Number of items at targeted PL 5
SI	0	0	0	3	0
LA	0	0	0	0	0
MA	0	3	3	0	3
SC	3	0	0	0	0
SS	0	0	0	0	0

Table 2.4.3.5. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Writing Narrative storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2 to 5
SI	1	0
LA	0	0
MA	0	0
SC	0	0
SS	0	0
IT (SIL, LoLA, LoSS)	0	1

Table 3.4.3.6. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Writing Expository storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4/5
SI	1	0	0	0
LA	0	0	0	0
MA	0	3	0	0
SC	0	0	0	0
SS	0	0	4	0
IT (SIL, LoLA, LoSS)	0	0	0	1

Table 4.7 Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Reading Narrative storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4	Number of items at targeted PL 5
SI	3	0	0	0	3
LA	0	0	0	0	0
MA	0	0	0	0	0
SC	0	3	3	0	0
SS	0	0	0	3	0

Table 2.4.3.8. Number of items by WIDA Standard and targeted proficiency level on Kindergarten ACCESS: Reading Expository storyline

WIDA Standard	Number of items at targeted PL 1	Number of items at targeted PL 2	Number of items at targeted PL 3	Number of items at targeted PL 4	Number of items at targeted PL 5
SI	3	0	0	3	0
LA	0	0	0	0	0
MA	0	0	0	0	0
SC	0	3	3	0	3
SS	0	0	0	0	0

3. Test Administration

3.1 Test Delivery

Administration of ACCESS Paper typically takes place between December and April of the academic year, with testing windows determined at the state level. The domain tests for grades 1–12 may be administered in any order. The test may be administered in several sessions within one day or over a series of days.

The Listening and Reading tests may be group- or individually administered. Students are administered the Listening and Reading test forms using paper test booklets, and students record their answers directly in the test booklets. For the Listening test, the audio stimuli are played aloud via an audio CD.

The Writing test may be group- or individually administered. Students are administered the Writing test via paper test booklets. Students record their responses directly in the test booklet.

The Speaking test is individually administered. Students listen to an audio recording and follow along in an accompanying test booklet. Each task also includes a model student response, which serves as an exemplar to the student and as a benchmark to the Test Administrator who scores the task. All audio stimuli are presented via audio CD.

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 Administrator, and Test Coordinators at both the district and school levels. 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 in the WIDA Secure Portal.

The training courses within the WIDA Secure Portal are 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 an online administration quiz.

The roles of the test administrator and technology coordinator are critical for the proper administration of the assessments as proper training and familiarity with ACCESS for ELLs administration requirements is key to the validity of the test and the appropriate interpretations of test scores.

3.2.1 Listening Test Administration

The ACCESS for ELLs Paper Listening test is media delivered. Listening test items are delivered via CD.

3.2.1.1 Listening Test Materials

Test materials include the following items:

- Test Administrator Script
- Student Test Booklet(s)
- Listening and Speaking Test CD (a separate CD for each grade-level cluster and tiered test form). In the rare event that a student requires a human reader as an accommodation, the Human Reader Accommodation Script is required to administer the Listening section individually for that student.
- At least one sharpened number 2 pencil for each student to mark responses
- Speakers
- A CD player or desktop/laptop computer (to play the CD)

3.2.1.2 Organization and Timing of the Listening Test

The Listening test is designed to take approximately 25 to 40 minutes, depending on the grade-level cluster and tier. The test administration time does not include time for convening students, taking attendance, distributing, and collecting test materials, explaining test directions, or completing practice items. The length of test items increases with students' language proficiency and grade level. For example, the Tier B/C Listening test takes longer to administer than the Tier A Listening test, and the Listening test for grades 9–12 may take slightly longer than the test for grades 4–5.

3.2.2 Reading Test Administration

The ACCESS for ELLs Reading test is completed within Student Test Booklets after a scripted introduction by the Test Administrator.

3.2.2.1 Reading Test Materials

- Reading test materials include the following items:
- Test Administrator Script
- Student Test Booklet(s)
- At least one sharpened number 2 pencil for each student to mark responses

3.2.2.2 Organization and Timing of the Reading Test

The Reading test is designed to take no more than 35 to 45 minutes. The test administration time does not include time for convening students, taking attendance, distributing, and collecting test materials, explaining test directions, or completing practice items.

3.2.3 Writing Test Administration

Students respond to a set of tasks, writing their responses in their Student Test Booklets.

3.2.3.1 Writing Test Materials

Writing test materials include the following items:

- Test Administrator Script
- Student Test Booklet(s)
- At least one sharpened number 2 pencil for each student to write responses
- Scratch paper

3.2.3.2 Organization and Timing of the Writing Test

There are three tasks (Parts A, B, and C) in each Tier (Tiers A and B/C) of the Writing test for all grade levels except Tier A for grade 1, which contains four tasks. For grade-level clusters 2, 3, 4–5, 6–8, and 9–12, the Tier A Writing tests have recommended guidelines for Parts A, B, and C of 15 minutes each, with up to 5 additional minutes for each part if needed for students to finish writing, for a total of 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.

3.2.4 Speaking Test Administration

The ACCESS for ELLs Speaking test is an individually administered test that standardizes test administration across students. Speaking test items are media delivered. Speaking test audio is provided on the same CD as the Listening test. The Speaking test provides ELs with the opportunity to demonstrate their academic English language proficiency in Speaking across the WIDA ELD Standards through a set of constructed response tasks. The Speaking test is tiered. Students will either take the Tier A form or the Tier B/C form; both are included in the same Speaking Test Booklet.

3.2.4.1 Audio Format of the Speaking Test

The Speaking test is multimodal. The student hears audio input and sees the input as text in the Speaking Test Booklet. This presentation format supports the student in understanding test input. Media delivery of the Speaking test means that an audio recording will guide the student through the Speaking test. The audio recording includes two voices: a model student and a Virtual Test Administrator.

Each task on the Speaking test is preceded by a model student task and response. The questions posed to the model student are at the same proficiency level as the tasks to which the student will respond, allowing the model student to demonstrate the expected language use at a given proficiency level. In most cases the model questions are designed to be parallel to, but not exactly the same as, the examinee questions. The model student also has an important function in scoring since the scoring scale is designed to evaluate student responses relative to the model student's response.

The Virtual Test Administrator guides the student through the test and asks the student questions designed to elicit language at targeted proficiency levels. While the Virtual Test Administrator will instruct and guide the student through the Speaking test, the administrator may also need to assist the student in navigating test materials (e.g., turning the page when prompted). The Speaking test includes standardized, built-in response time for every task. The amount of time varies according to the grade-level cluster, tier, and proficiency level of the task and ranges from 15 to 50 seconds in grades 1–3 and from 15 to 45 seconds in grades 4–12. Students may not require the entire time allotted. After the response time has ended, the test audio will automatically continue to the next Speaking task.

3.2.4.2 Speaking Test Materials

Speaking test materials include the following items:

- Test Administrator Script
- Speaking Test Booklet (contains test graphics and prompts)
- Student Test Booklet (contains Speaking test scoring sheet and scoring scale)
- Listening and Speaking test CD (a separate CD for each grade-level cluster and tiered test form). In the rare event that a student requires a human reader as an accommodation, the Recording Script is required to administer the Speaking section.
- A CD player or desktop/laptop computer (to play the CD)
- Speakers

3.2.4.3 Organization and Timing of the Speaking Test

Speaking tasks on the Speaking test are contained within three parts: A, B, and C. As in other domains of ACCESS for ELLs, tasks on the Speaking test are grouped thematically. Each part addresses one or more of the WIDA ELD Standards and contains two tasks. In all, the Speaking test contains six individual tasks across the three parts. Each task is associated with a proficiency level (1, 3, or 5) and includes one or two questions to which the student responds. Student questions are indicated by a blue speech bubble in the test booklet.

The Speaking test is designed to take approximately 15 to 35 minutes per student, but the actual time will depend on the grade-level cluster and tier of the test administered. Note that the approximate test administration time does not include setting up the test session or explaining test directions. An additional 10 minutes should be allocated to set up the Speaking test.

3.2.5 Test Administrator Training

To prepare individuals to serve as Test Administrators, Test Administrator training for ACCESS Paper is conducted through online training modules hosted within the WIDA Secure Portal. Three certifications are offered to participants: a group test administration certification pertaining to the Listening, Reading, and Writing portions of ACCESS; a certification for the Speaking test; and a certification for Kindergarten ACCESS. To receive any of the three certifications, participants must complete the relevant online course and pass a qualifying exam after completing the course.

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 (AMS) and the DRC 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 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, housed in the SEA Secure Portal, and the Memorandum of Understanding (MOU)s 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 programs (IEPs) or 504 Plans. These supports are intended to increase accessibility to the assessments for all ELs. Please see the Accessibility and Accommodations Manual, available on the public WIDA website, for more details. 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 *Test and Item Design Plan for ACCESS for ELLs Paper and WIDA Screener Paper,* housed 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 test takers. 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, 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 Paper and Kindergarten ACCESS to address their accessibility needs. Audio aids, color overlay, highlighters, colored pencils or crayons, line guide or tracking tool, low-vision aids or magnification devices, sticky notes, and scratch paper are the universal tools used in the ACCESS Paper administration.

3.3.2 Support Provided to ELLs 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 ELs with disabilities who have 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 *WIDA 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.

Universal tools are also available to all students taking ACCESS for ELLs. All accessibility features are available to all ELs during testing; specific designation is not required prior to testing to make them available to the student. Features available during paper-based test administration include the following:

- Audio amplification device (provided by student)
- Highlighter, colored pencils, or crayons
- Place marker (blank)
- Low-vision aids or magnification device
- Color overlay
- Equipment or technology that the student uses for other tests and schoolwork, e.g., adapted pencil (altered size or grip), slant board, wedge, etc.
- Scratch/blank paper (submit with test or dispose 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 test developers at CAL.

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

4. Scoring Procedures

4.1 Multiple Choice Scoring: Listening and Reading

Listening and Reading items are scored dichotomously, as correct or incorrect. Students mark their answers directly in their test booklets, and each page is scanned into an electronic database. Scale scores for each domain are calculated based on the items that are administered to the student and the number 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 Writing

Trained raters score students' responses to the performance-based tasks in the domain of Writing. 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 applications 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 undergo thorough training and qualifying. Training is task-specific to ensure that raters understand the nuances of each unique Writing task. DRC selects team leaders based on their prior performance as raters and for their leadership skills. They are assigned to small groups of raters, typically 7 to 10 raters on each team. The team leaders are responsible for monitoring the performance of their team members and providing ongoing feedback to support accurate scoring. DRC promotes scoring directors, who earn their positions by demonstrating quality work as raters and as team leaders on previous projects, from within. Scoring directors are responsible for a specific set of tasks within a single domain. The scoring directors train and oversee the teams of raters assigned to these tasks. What follows are general scoring procedures utilized by DRC.

Preparing rater training materials for Writing tasks

CAL test development staff produce materials that DRC uses to train their raters to score ACCESS Writing responses. CAL test development staff members who are trained on the Writing Scoring Scale ("Expert Raters") prepared these rater training materials for Writing tasks when they were originally developed for ACCESS Online. Given that the ACCESS Paper Writing test is not refreshed annually, the rater training materials carry over from year to year.

To prepare the Writing rater training materials, the Expert Rater began 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 reviewed documented criteria for anchor responses and score explanations.

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

Following the Expert Rater's initial review of responses, the Writing Test Development Manager (TD Manager) reviewed the responses selected. The TD Manager confirmed or revised the scores, recording notes and feedback, and finalized 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 tasks with primarily handwritten responses, the handwriting must also be generally legible to facilitate internalization of the linguistic characteristics by raters.

Once anchor responses were finalized, the Expert Rater wrote 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 reviewed the score explanations to check that they met the required criteria. The TD Manager also selected 20 responses from the initial review to be used as training samples, and reviewed and revised any accompanying score notes, as necessary. The 20 training samples were 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 reviewed all notes from the anchor and training sample selection process and, when necessary, compiled any task-specific scoring guidance to be used by raters.

The anchors, explanations, training samples, training sample notes, and any task-specific scoring guidance were then provided to WIDA for review. CAL staff updated the materials as requested by WIDA and delivered 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% 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 or 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.
- 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 20% of the responses were scored a second time. The responses comprising the 20% denoted as 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—or the entirety—of 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, the 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 five 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.
- **Non score 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 after 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 strive 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 characteristics of DRC's remote scoring practices are as 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.
- **Rater training.** 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. 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. 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.
- **Chat tool.** 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.

- **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).

Figure 4.3 shows the Writing Scoring Scale.

Figure 4.3. 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. They are defined as follows:

- **Nonscorable:** The response is blank; consists only of verbatim copied text; consists only of text that is completely off task; 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 three (or four, for grade 1 Tier A) Writing tasks. One rater assigns a score to each response. To calculate a student's total raw score, the scores that the raters assigned are converted to whole numbers ranging from 0 to 9, as shown in Table 4.3.

Table 4.3. 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

On Tier A tests, for all grade-level clusters except for grade 1, the scores from the three tasks are added to calculate a total raw score, which can range from 0 to 27. For the grade 1 Tier A test, there are four Writing tasks. The first two of these tasks use a modified version of the scoring scale and have score ranges of 0 to 1 and 0 to 3, respectively. The third and fourth tasks use the full scoring scale from 0 to 9; additionally, the last task is weighted as 3. Therefore, the possible final raw scores for grade 1 Tier A range from 0 to 40.

On Tier B/C tests for all grade-level clusters, results from the different tasks are given different weights. These weights are specified to reflect the intended amount of time that a student should spend on each task. The first task is given a weight of 1, the second task is given a weight of 2, and the third task is given a weight of 3. Thus, for example, a student with raw scores of 5, 6, and 7 on the three tasks would have a total raw score of 38 ([1 * 5] + [2 * 6] + [3 * 7]), while a student with raw scores of 7, 6, and 5 on the three tasks would have a total raw score of 34 ([1 * 7] + [2 * 6] + [3 * 5]). Raw scores on the Tier B/C tests can range from 0 to 54.

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 test is scored using a scoring scale that is designed to evaluate student responses relative to the model student's response. (See Section 2.2.4 for more information about the role of the model student in the design of the Speaking tasks.) As part of test administration, the test administrators hear the model student response before each student response, which supports them in assigning an appropriate score relative to the model response. Speaking responses are immediately scored by the administrator while the test is administered. After listening to the student's responses, the administrator assigns a score.

The Speaking Test is the only portion of ACCESS Paper that is scored locally. Test administrators must complete the relevant virtual ACCESS Paper Test Administrator training module for the Speaking test and pass the accompanying quiz (either Grades 1–5 or Grades 6–12). The training focuses on developing the test administrators' ability to score the test reliably. Separate training materials are available that address test administration and monitoring procedures. To help ensure that test administrators reliably score the test, they are trained on the Speaking Scoring Scale. Training materials are available for each grade-level cluster, and raters listen to anchor samples and view score justifications that provide detailed explanations for scores based on the scoring scale. Practice samples are also available so that raters can practice assigning scores. Although the ACCESS Paper Test Administrator training module for the Speaking test was created by the WIDA instructional design team, CAL test development staff provided the anchor samples, score justifications, and practice samples to be used in the training module. These materials were identified and created in a manner analogous to the Writing rater training materials, as described in Section 4.2. The course includes both required training material for each grade-level cluster as well as optional training material. Raters are required to complete training sections for each grade-level cluster they will administer and score. However, if raters score more than three grade-level clusters, they may complete rater training for only three. The quizzes include 12 Speaking rating tasks in which raters listen to and assign a score to a task response. The pass rate for the quiz is 80% correct.

The Speaking Scoring Scale defines five score points: *Exemplary, Strong, Adequate, Attempted,* and *No Response (in English).* The No Response score point only applies if the examinee refuses to respond, or if the examinee responds in a language other than English.

These score points are applied based on the proficiency level expectations of each task, that is, the level of language proficiency that each task is designed to elicit. These expectations are exemplified by the model student response (see Section 2.2.4). In this way, the model response serves as a scoring benchmark. Raters listen to the model response and score test takers' responses relative to the model. A score of *Exemplary* means that the student response demonstrates English language use that is equal to or beyond the English language use illustrated by the model student's response.

Table 4.4.1 shows the Speaking Scoring Scale.

Table 4.4.1. WIDA Speaking Scoring Scale

Score Point	Response Characteristics
Exemplary use of oral language to provide an elaborated response	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	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	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	Language use does not support an adequate response Comprehensibility may be compromised Word choice may not be fully adequate
No response (in English)	Does not respond (in English)

The Speaking Scoring Scale includes descriptors for overall language use, response sophistication, language delivery, and word choice. As stated above, the scale is applied relative to the proficiency level demands of the task. 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.

To calculate a raw score for the Speaking test, the five score points are converted to whole numbers, as shown in Table 4.4.2. To calculate a total raw score, the raw scores for each task are added together; additionally, 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 the B/C test would easily achieve a score of *Adequate and Above* on these tasks. Thus, on the Tier A test, scores can range from 0 to 18; and on the Tier B/C test, from 6 to 30.

Table 4.4.2. Rating to raw score conversion (Speaking)

Score	Raw Score
No Response (in English)	0
Attempted	1
Adequate/Adequate and Above	2
Strong	3
Exemplary	4

Speaking tasks are scored using the WIDA Speaking Scoring Scale. The Speaking Scoring Scale is distinct from the WIDA Speaking Rubric, which is a tool for classroom use and score interpretation. The Speaking Scoring Scale was designed specifically for test scoring use and is not intended for classroom purposes.

5. Summary of Score Reports

5.1 Individual Student Report

Score reports (district, school, and student level reports) are made available in the 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 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. Communication about student score reports and resources that districts use to support interpretation is a local decision. 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 assessment results is determined locally.

Individual student reports are available in various languages in WIDA AMS and alternate format score reports (i.e., braille or large print) 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 Scores and Reports and Family Engagement pages on the WIDA website.) WIDA also provides a post-testing Q & A webinar about score interpretation, located in 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 explanation when describing student performance. WIDA has convened focus groups to gain a better 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 5.1) contains detailed information about the performance of a single student in grades K–12. Its primary users are students, parents/guardians, teachers, and other educators. It provides information about the language needed to access content and succeed in school, one indicator of a student's English language proficiency.

Figure 5.1. Individual Student Report

WIDA UNIVERSITY OF WISCONSIN-MADISON | **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) 1 2 3 4 5 6	Scale Score (Possible 100-600) and Confidence Band See Interpretive Guide for Score Reports for definitions 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: • Identify main topics in discussions • Categorize or sequence information presented orally using pictures or objects • Follow short oral directions with the help of pictures • Sort 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: • Share about what, when, or where something happened • Compare objects, people, pictures, and events • Describe steps in cycles or processes • Express opinions
Reading	2	understand written language related to specific familiar topics in school and can participate in class discussions, for example: • Identify main ideas in written information • Identify main actors and events, in stories and simple texts with pictures or graphs • Sequence pictures, events or steps in processes • Distinguish between claim and evidence statements
Writing	4	communicate in writing in English using language related to specific topics in school, for example: • Produce papers describing specific ideas or concepts • Narrate stories with details of people, events, and situations • Create explanatory text that includes details or examples • Provide 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 a disability, WIDA provides states methods to compute alternative composite scores based on their available domain scores upon request (Sahakyan, N., (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. A scale score represents a student's performance that has been put on a standardized scale. A student's 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. 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. ELs 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 associated confidence band. The confidence band for each domain and composite 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 they 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 their level of proficiency. For example, as shown in Figure 5.1, this student received a proficiency level score of 2 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 primary 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.

Both Student Roster Reports and Frequency Reports can be accessed in WIDA AMS.