Annual Technical Report
ACCESS for ELLs
Online English Language Proficiency Test
Series 601, 2022–2023 Administration

Annual Technical Report No. 19A

Prepared by

Center for Applied Linguistics

Language Assessment Division
Psychometrics and Quantitative Research Team

Nov 2023

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 19th annual technical report on the ACCESS for ELLs English Language Proficiency Test and the seventh report on the ACCESS for ELLs assessment given in Online format.

This technical report is produced as a service to members and potential members of the WIDA Consortium and to support states' submissions for U.S. Department of Education English language proficiency assessment peer review. The technical information herein is intended for use by those who have technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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 2022–2023 ACCESS assessment.

ACCESS for ELLs is intended to assess reliably and validly 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 Online Series 601 was administered in school year 2022–2023 in 35 states, the Bureau of Indian Education, the Department of Defense Education Activity, the Northern Mariana Islands, the Virgin Islands, and the District of Columbia for a total of 40 state entities (henceforth "states").

The Series 601 Online data set used in this report included the results of 1,838,523 students as of September 2022. The final number of students who participated in the Series 601 Online ACCESS tests is 1,998,137. The grade with the greatest number of students represented in this report was Grade 2 with 224,672 students, while the grade with the smallest number was Grade 12, with 79,380 students. Of the participating WIDA states, the state with the largest population of EL students was Illinois, with 224,014 students, while the state with the fewest was the U.S. Virgin Islands, with 1,042 students.

Based on a comparison with prior years' numbers of participating students, there is a 6% increase in the student population that participated in ACCESS Series 601 testing than ACCESS Series 503 testing.

ACCESS for ELLs Series 601 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 19A) provides technical information pertaining to ACCESS for ELLs Series 601 Online. A second report (WIDA ACCESS Technical Report 19B) provides technical information for the ACCESS for ELLs Series 601 Paper assessment, including the Kindergarten assessment.

Part 1

Purpose, Design, Implementation

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 language learners (ELLs) in Grades K–12 in the 41 U.S. states, territories, and federal agencies in the WIDA Consortium, first in the English Language Proficiency Standards (Gottlieb, 2004; WIDA 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 ELLs. ACCESS may thus be described as a standards-based English language proficiency test designed to measure the social and academic language proficiency of ELLs in English. It assesses social and instructional English as well as the academic language associated with language arts, mathematics, science, and social studies, within the school context, across the four language domains (Listening, Reading, Writing, and Speaking).

Other purposes of ACCESS include

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

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

1.2 The WIDA Standards

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

- *Standard 1*: ELLs communicate in English for **Social and Instructional** purposes within the school setting.
- *Standard 2*: ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Language Arts**.
- *Standard 3*: ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Mathematics**.
- *Standard 4*: ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Science**.
- *Standard 5*: ELLs communicate information, ideas, and concepts necessary for academic success in the content area of **Social Studies**.

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

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

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

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

- Language of Language Arts (LoLA) and Language of Social Studies (LoSS): LS (Speaking and Writing)

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

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

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

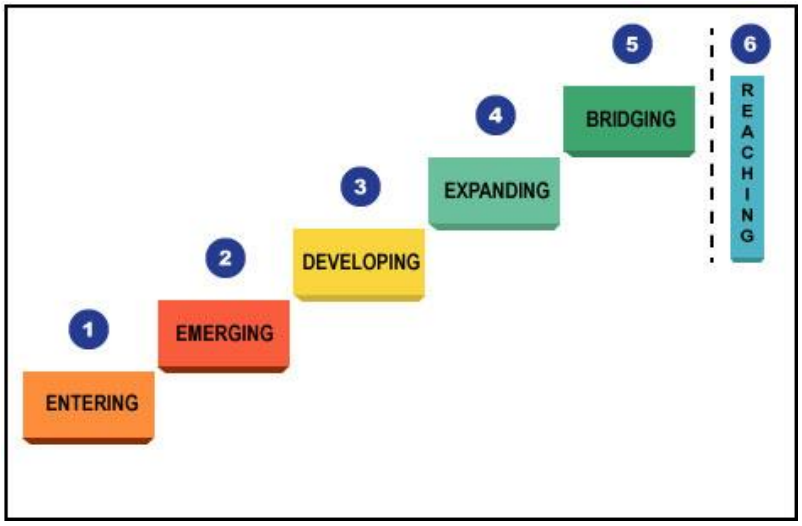


Figure 1. The language proficiency levels of the WIDA ELD Standards

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

First, they appear in the **performance definitions**. The performance definitions describe the stages of language acquisition, providing details about the language that students can comprehend and produce at each proficiency level. The performance definitions are based on three criteria: (1) vocabulary usage at the word/phrase level; (2) language forms and conventions at the sentence level; and (3) linguistic complexity at the discourse level. Vocabulary usage refers to students' increasing comprehension and production of the technical language required for success in the academic content areas. Language forms and conventions refers to the increasing development of phonological, syntactic, and semantic understanding in receptive skills or control of usage in productive language skills. Linguistic complexity refers to students' demonstration of oral interaction or writing of increasing quantity and variety.

Second, language proficiency levels are represented through connections to the accompanying **Model Performance Indicators** (MPIs). The MPIs provide a model of the expectations for 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 proficiency levels 1 through 5 for a given WIDA Standard, grade-level cluster, domain, and topic is called a strand. These MPIs together describe a logical progression and accumulation of skills on the path from the lowest level of English language proficiency to full English language proficiency for

academic success. The final level, PL 6: *Reaching*, represents the end of the continuum rather than another level of language proficiency.

Each MPI has a tripartite structure, consisting of a language function, a content stem, and support. The MPIs used on ACCESS can be taken directly from the WIDA English Language Proficiency Standards (WIDA 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 purposes of 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 Consortium, 2021d). 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 Online is as follows: 1, 2–3, 4–5, 6–8, and 9–12. Note that the Kindergarten (K) form is not administered online and thus is not covered in this report.

1.6 Tiers

ACCESS is designed so that test paths or forms are appropriate to the proficiency level of individual students across the wide range of proficiencies described in the WIDA ELD Standards (Figure **2**). Tests must be at the appropriate difficulty level for each individual student to facilitate valid and reliable interpretations of scores. While the grade-level cluster structure is a design feature intended to ensure that the language expectations are developmentally appropriate for students in different age ranges, within each grade-level cluster, students display a range of abilities. Test items and tasks designed for Entering (PL 1) or Emerging (PL 2) students to demonstrate accomplishment of the MPIs at their proficiency level will not allow Expanding (PL 4) or Bridging (PL 5) students to demonstrate the full extent of their language proficiency. Likewise, items and tasks that allow Expanding (PL 4) and Bridging (PL 5) students to demonstrate accomplishment of the MPIs at their level would be far too challenging for Entering (PL 1) or Emerging (PL 2) students. Items that are far too easy for students may be boring and lead to inattentiveness; items that are far too difficult for students may be frustrating and discourage them from performing their best. But more importantly, items that are too easy or too hard for a student add very little to the accuracy or quality of the measurement of that student's language proficiency.

In the Listening and Reading multistage adaptive tests, students are routed to folders that vary in difficulty, designated as A, B, or C level folders.[1] Tier A folders are intended for students at beginning levels of English language proficiency (PLs 1–3), Tier B folders for students at intermediate levels (PLs 2–4), and Tier C folders for students at more advanced proficiency levels (PLs 3–5). In the domain of Writing, the test forms are designated as either Tier A, which includes tasks written to elicit language up to PL 3, or Tier B/C, which includes tasks written to

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

elicit language up to PL 4 or PL 5. In the domain of Speaking, test forms are designed so that students at very beginning levels of proficiency take a pre-A form, which is designed to elicit language at PL 1; students at early levels of proficiency take the Tier A form, with tasks designed to elicit language at PL 1 and PL 3; and more proficient students take the Tier B/C form, with tasks designed to elicit language at PL 3 and PL 5.

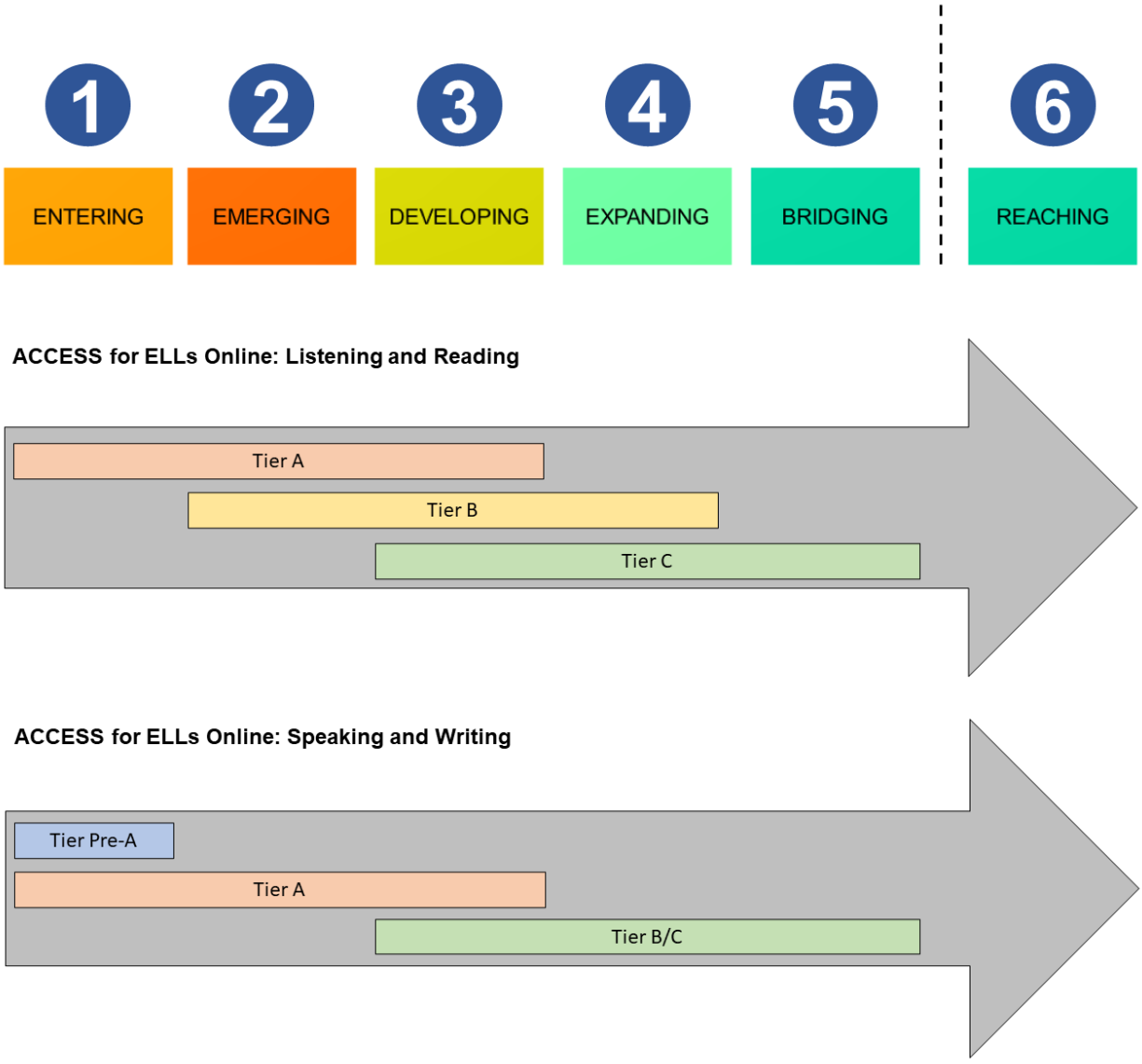


Figure 2. Tiers and proficiency levels

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 purposes of 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 ACCESS Test Practice and Sample Items page on WIDA's website, https://wida.wisc.edu/assess/access/preparing-students/practice.

When the task models for ACCESS for ELLs Online were first developed, CAL and WIDA addressed issues of fairness by ensuring that principles of Universal Design of Assessments (UDA) (National Center on Educational Outcomes, 2021) were adhered to in this design phase. Therefore, CAL, WIDA, and Data Recognition Corporation (DRC) collaborated to design the item and task layout on the screen to be maximally readable/legible with sufficient whitespace, to be accessed intuitively by students, to be accompanied by instructions and practice items to allow students to become accustomed to the test interface, and to contain universal accessibility tools (magnifier, line guide) as well as tools for accommodation (such as control of test audio and extended response time for the Speaking test). The ways in which 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 below.

2.1.1 Listening Items

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

Most items on the operational Listening test are traditional multiple choice, though some operational items and some items embedded for field testing purposes may involve enhanced item presentations, including hot spot items (i.e., the student clicks on an area of the screen to

respond) and drag-and-drop items (i.e., the student drags an image/text to a specified screen area to respond).

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

Figure 3. Multiple choice item layout for the ACCESS for ELLs Online Listening test

For hot spot items, students see a large response area. The response area may be an image, a paragraph of text, or some combination of images and text, such as a timeline or a webpage. The answer choices may be pictures or text and are embedded in the response area inside blue boxes. Students answer the question by clicking on one of the boxes in the response area. Each answer choice changes color when selected. Students can change their answers by clicking on a different

blue box or by clicking on the reset eraser button, which clears the original response, and clicking on a different blue box. A screenshot of a sample hot spot item is provided in Figure 4.

Figure 4. Layout of a Hot Spot item for the ACCESS for ELLs Online Listening test

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

select one object to drag into the single blue box within the response area. A screenshot of a sample drag and drop item is provided in Figure 5.

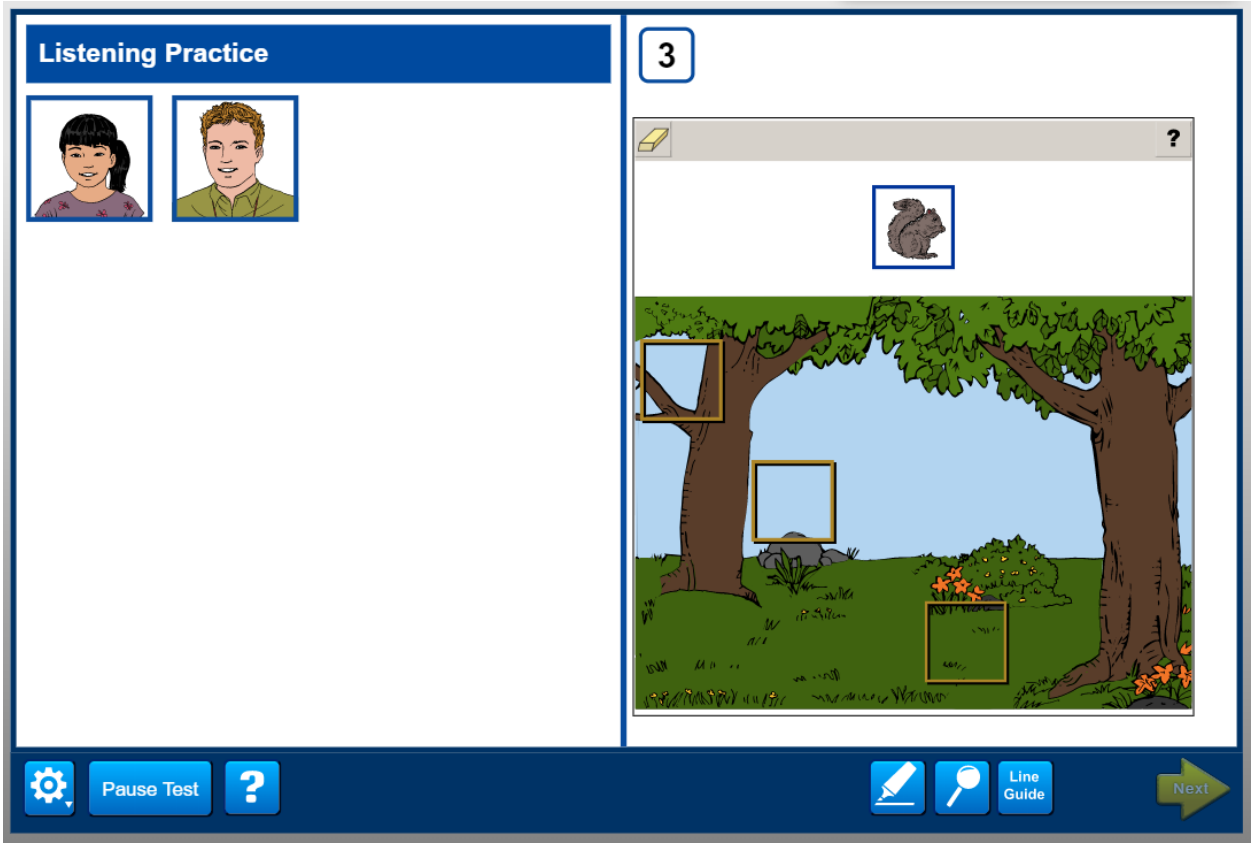


Figure 5. Layout of a drag and drop item for the ACCESS for ELLs Online Listening test

The number of enhanced items on the Listening test is not specified in the test or item specifications, so the appearance of enhanced items on the test is emergent from the content. In other words, if the content of a given item lends itself well to an enhanced item type, then it is operationalized as such.

Each item on the Listening test targets the language of one of the five WIDA ELD Standards and tests a student's ability to process language at one of the five fully delineated proficiency levels.[2]

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

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

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

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

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

2.1.2 Reading Items

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

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

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

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

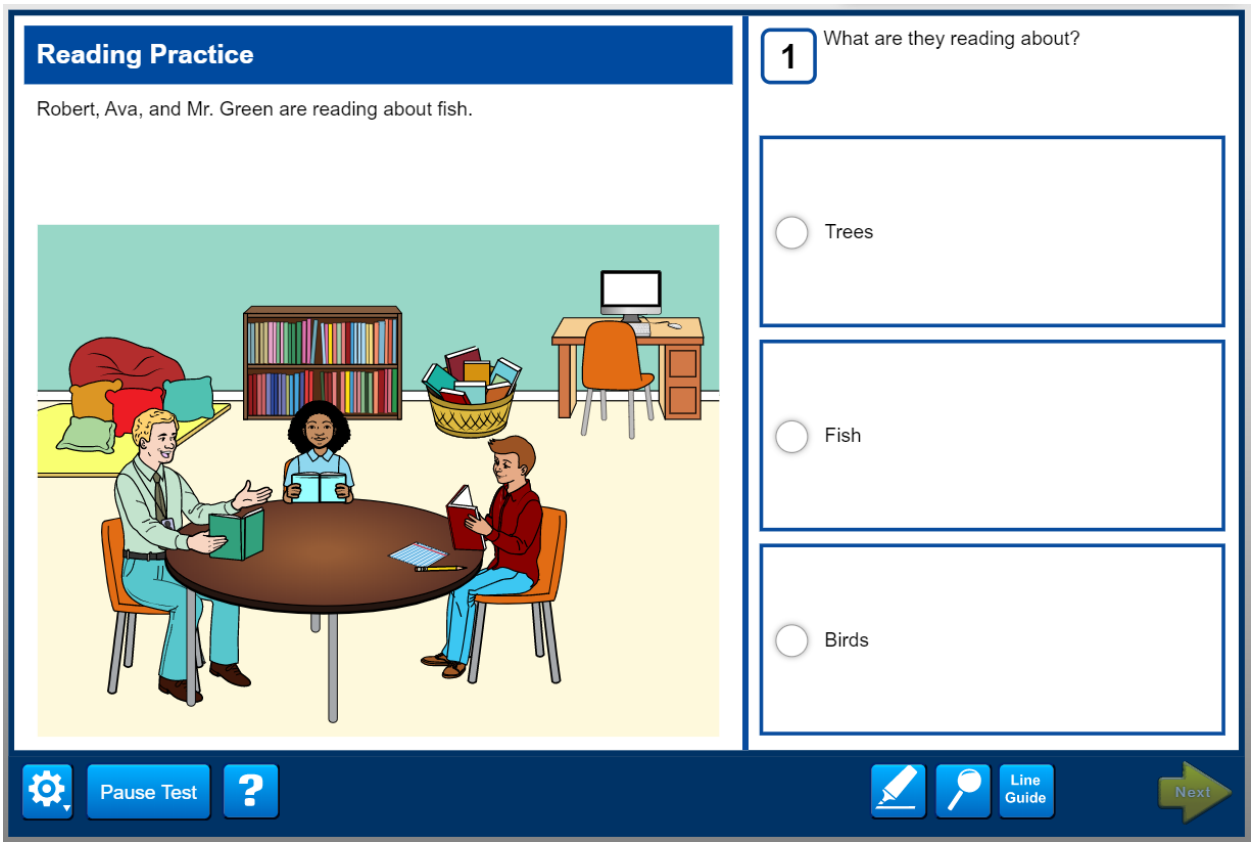


Figure 6. Multiple choice item layout for the ACCESS for ELLs Online Reading test

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

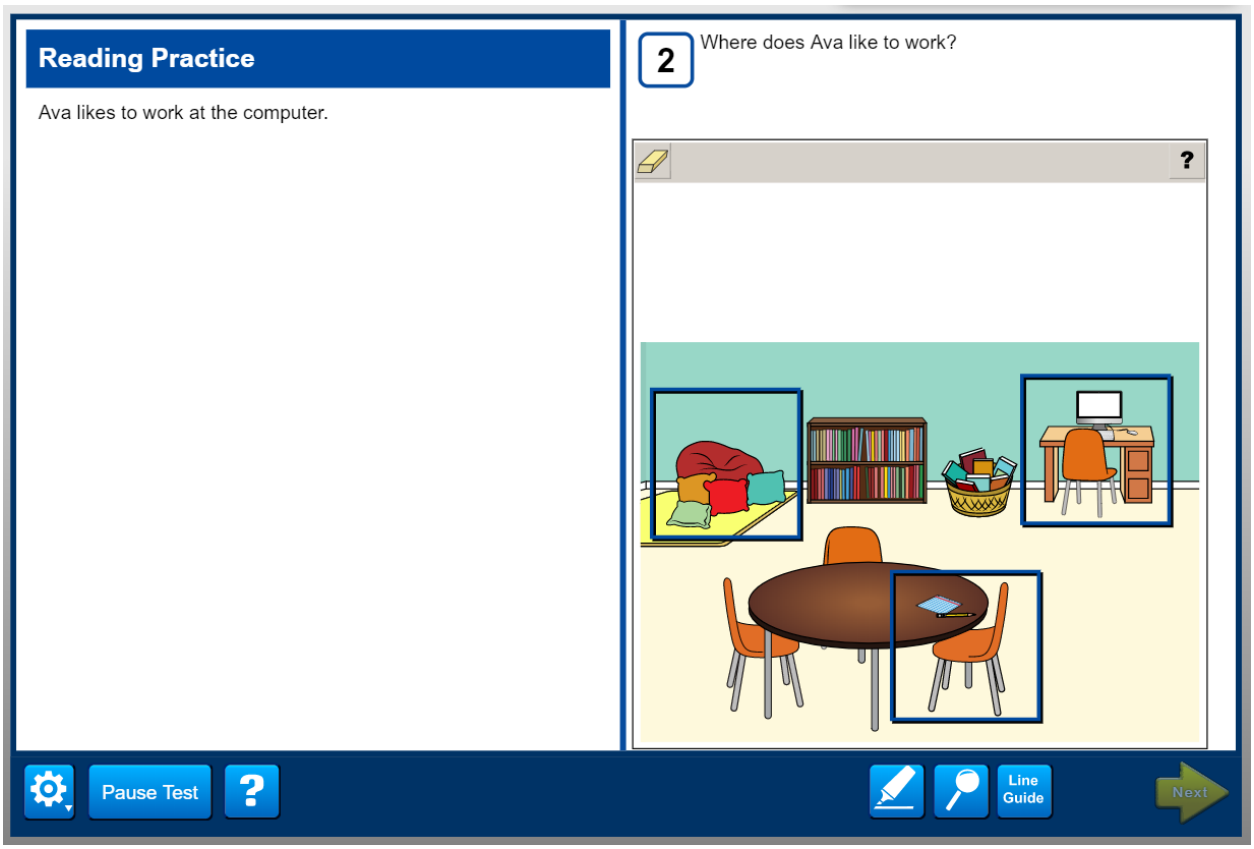


Figure 7. Layout of a hot spot item for the ACCESS for ELLs Online Reading test

Figure 8. Layout of a drag and drop item for the ACCESS for ELLs Online Reading test

The number of enhanced items on the Reading test is not specified in the test or item specifications, so the appearance of enhanced items on the test is emergent from the content. In other words, if the content of a given item lends itself well to an enhanced item type, then it is operationalized as such.

Items have one key and either two or three distractors, depending upon the grade-level cluster and the targeted proficiency level. For Grades 1 and 2–3, all items have a key and two distractors. For Grades 4–5, 6–8, and 9–12, items targeting PLs 1 and 2 have a key and two distractors, and items targeting PLs 3, 4, and 5 have a key and three distractors.

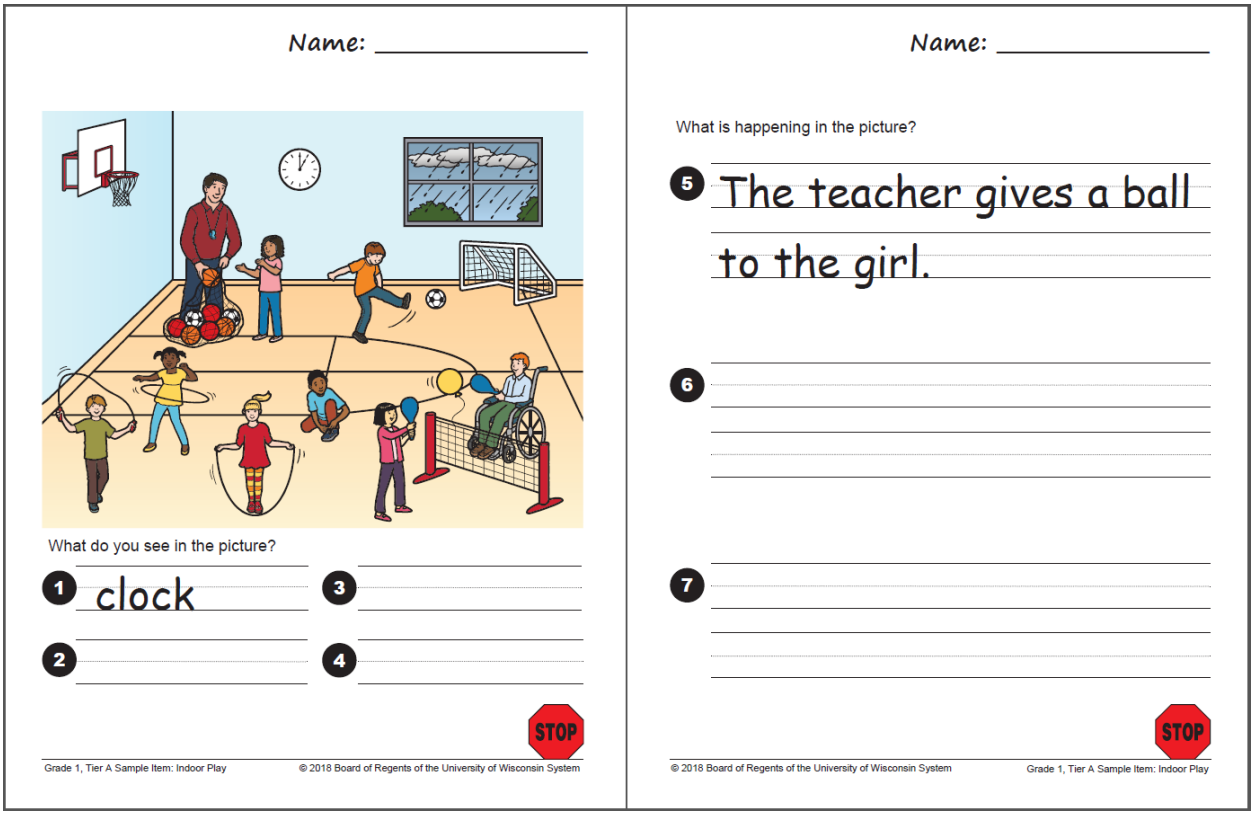
2.1.3 Writing Tasks

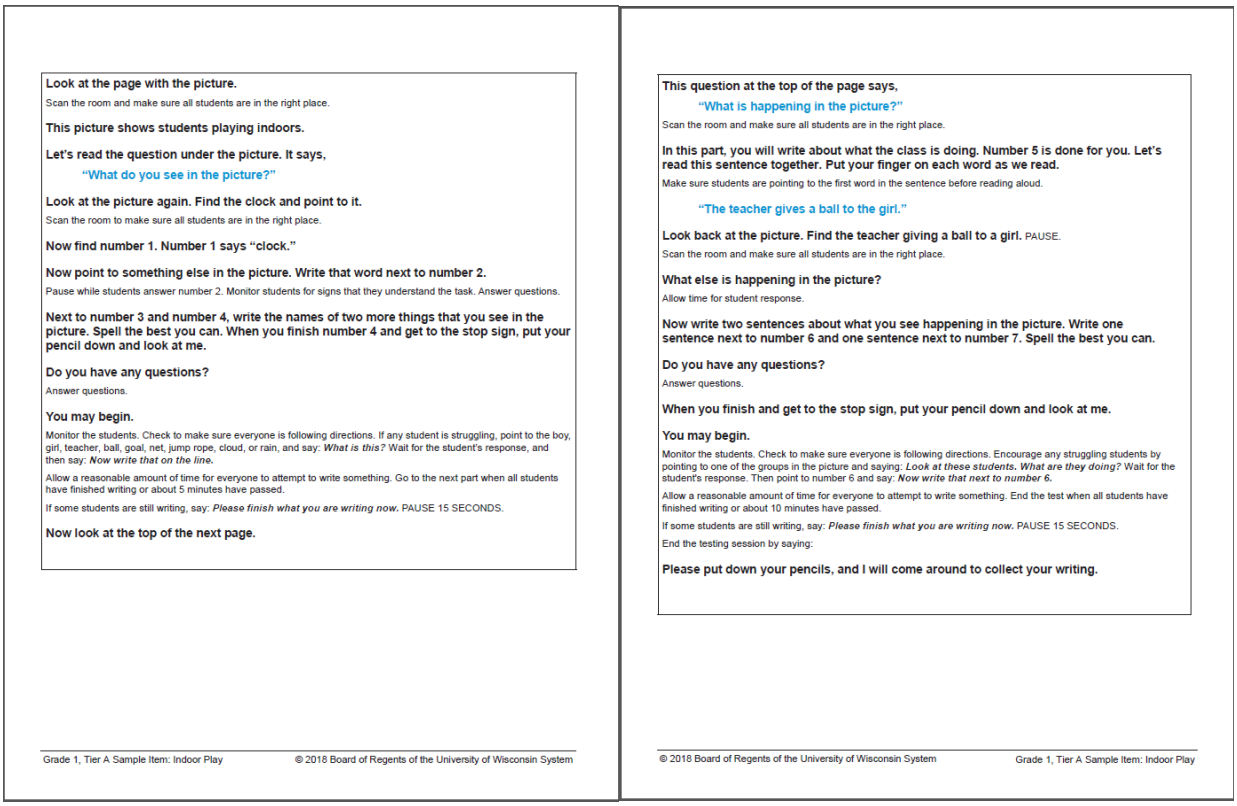
Writing tasks are designed to elicit language corresponding to one or more of the WIDA ELD Standards. Tasks appearing on the Tier A test form are designed to give students the opportunity to produce writing samples that fulfill linguistic expectations up to PL 3. DRC raters score students' written responses to these tasks using the entire breadth of the scoring scale. (For more information about scoring the Writing test, see Section 2.2.3 below.) Therefore, students may

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

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

Figure 9. Example test booklet for the ACCESS for ELLs Online Writing test, Grades 1–3





Look at the page with the picture.

Scan the room and make sure all students are in the right place.

This picture shows students playing indoors.

Let's read the question under the picture. It says,

"What do you see in the picture?"

Look at the picture again. Find the clock and point to it.

Scan the room to make sure all students are in the right place.

Now find number 1. Number 1 says "clock."

Now point to something else in the picture. Write that word next to number 2.

Pause while students answer number 2. Monitor students for signs that they understand the task. Answer questions.

Next to number 3 and number 4, write the names of two more things that you see in the picture. Spell the best you can. When you finish number 4 and get to the stop sign, put your pencil down and look at me.

Do you have any questions?

Answer questions.

You may begin.

Monitor the students. Check to make sure everyone is following directions. If any student is struggling, point to the boy, girl, teacher, ball, goal, net, jump rope, cloud, or rain, and say: ***What is this?*** Wait for the student's response, and then say: ***Now write that on the line.***

Allow a reasonable amount of time for everyone to attempt to write something. Go to the next part when all students have finished writing or about 5 minutes have passed.

If some students are still writing, say: ***Please finish what you are writing now.*** PAUSE 15 SECONDS.

Now look at the top of the next page.

Grade 1, Tier A Sample Item: Indoor Play © 2018 Board of Regents of the University of Wisconsin System

This question at the top of the page says,

"What is happening in the picture?"

Scan the room and make sure all students are in the right place.

In this part, you will write about what the class is doing. Number 5 is done for you. Let's read this sentence together. Put your finger on each word as we read.

Make sure students are pointing to the first word in the sentence before reading aloud.

"The teacher gives a ball to the girl."

Look back at the picture. Find the teacher giving a ball to a girl. PAUSE.

Scan the room and make sure all students are in the right place.

What else is happening in the picture?

Allow time for student response.

Now write two sentences about what you see happening in the picture. Write one sentence next to number 6 and one sentence next to number 7. Spell the best you can.

Do you have any questions?

Answer questions.

When you finish and get to the stop sign, put your pencil down and look at me.

You may begin.

Monitor the students. Check to make sure everyone is following directions. Encourage any struggling students by pointing to one of the groups in the picture and saying: ***Look at these students. What are they doing?*** Wait for the student's response. Then point to number 6 and say: ***Now write that next to number 6.***

Allow a reasonable amount of time for everyone to attempt to write something. End the test when all students have finished writing or about 10 minutes have passed.

If some students are still writing, say: ***Please finish what you are writing now.*** PAUSE 15 SECONDS.

End the testing session by saying:

Please put down your pencils, and I will come around to collect your writing.

© 2018 Board of Regents of the University of Wisconsin System Grade 1, Tier A Sample Item: Indoor Play

Figure 10. Example script for the ACCESS for ELLs Online Writing test, Grades 1–3

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

Figure 11. Example orientation screen for the ACCESS for ELLs Online Writing test, Grades 4–12, both tiers

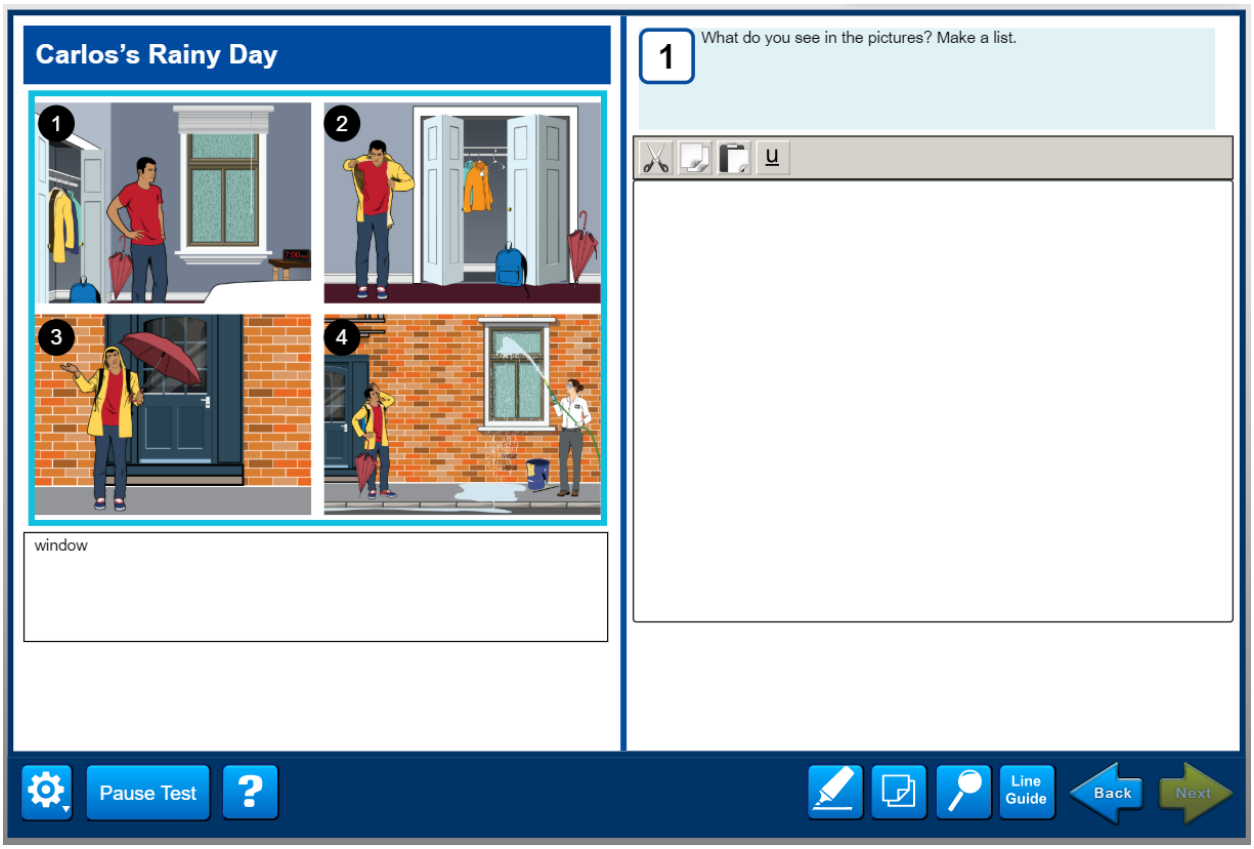


Figure 12. Example layout for the ACCESS for ELLs Online Writing test, Grades 4–12, Tier A

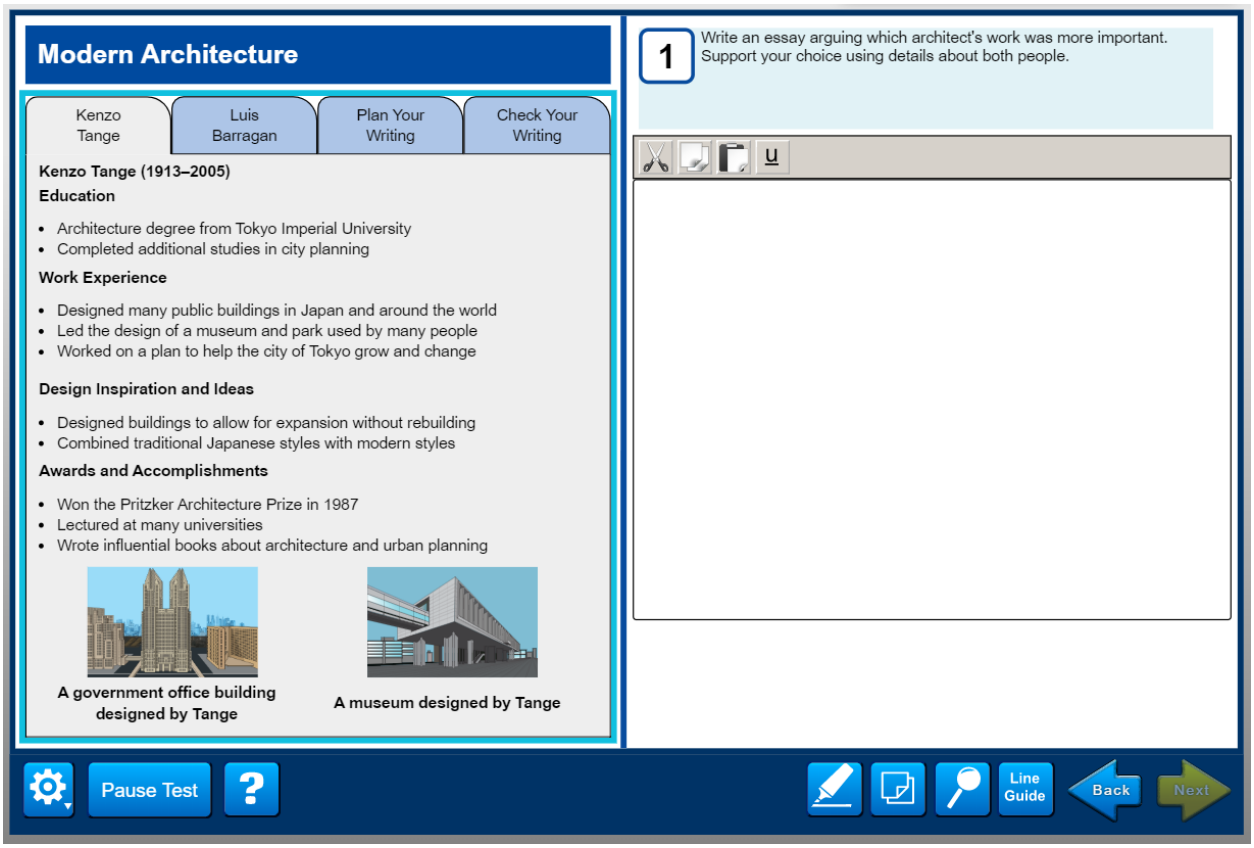


Figure 13. Example layout for the ACCESS for ELLs Online Writing test, Grades 4–12, Tier B/C

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

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

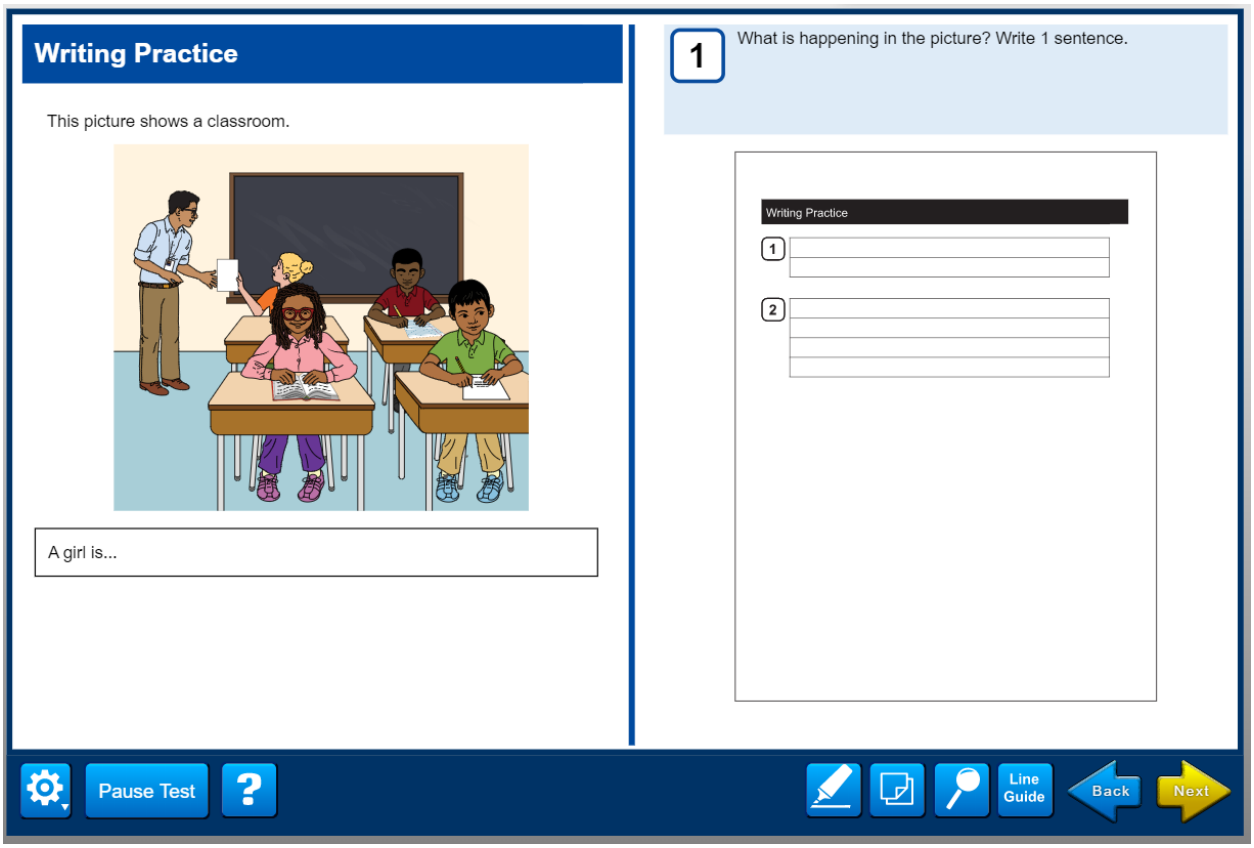


Figure 14. Example layout for an ACCESS for ELLs Online Tier A writing task with the handwritten (HW) response mode

2.1.4 Speaking Tasks

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

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

Students navigate through the Speaking test independently and at their own pace. They must listen to all audio on a screen before the test allows them to advance to the next screen. Most students can only listen to the audio stimuli once, although students with a specific

accommodation related to audio stimuli may listen to the audio as many times as they wish. The amount of time that students are allowed for recording their responses varies by grade-level cluster and the target proficiency level of the task; tasks targeting a higher proficiency level are permitted more recording time.[3] The amount and complexity of task input varies by grade-level cluster and task level. The purpose of the input is to provide academic content for students to draw on in their responses.

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

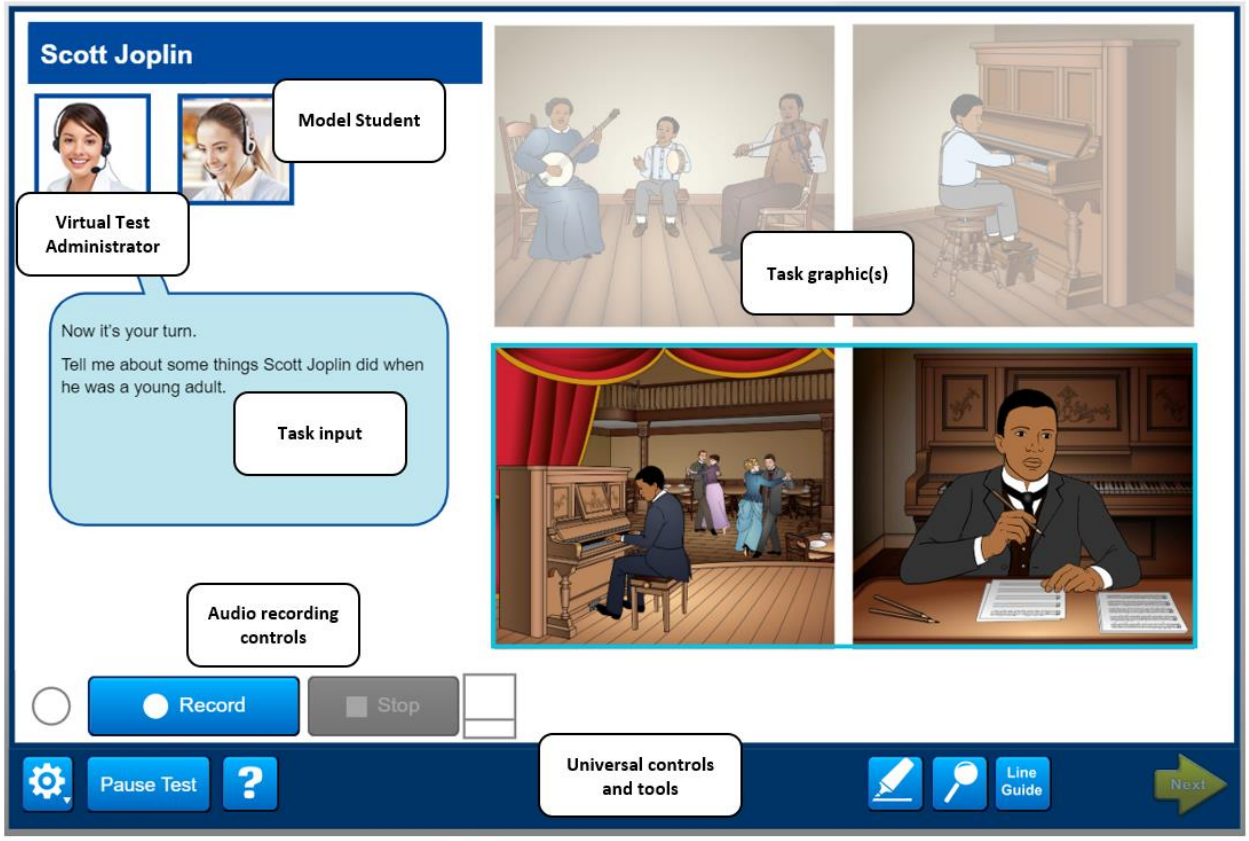


Figure 15. Visualization of the Speaking test screen layout

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

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

All Speaking tasks for a given grade cluster and WIDA Standard are designed in terms of *panels*; a panel is a thematically related set of three tasks, targeting the elicitation of PL 1, PL 3, and PL 5 language. When the tasks are field tested, the panels are split out into folders, with each folder containing one or two tasks. Tier Pre-A folders contain a single task targeting PL 1; Tier A folders contain two tasks targeting PL 1 and PL 3; and Tier C folders contain two tasks targeting PLs 3 and 5. For a given pair of Tier A and Tier C folders based on a single panel, the PL 3 task is identical in both folders (see Figure 7 in Section 2.2.4 for an illustration).

2.2 Test Design

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

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

2.2.1 Listening

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

type (MC – multiple choice; DD – drag-and-drop; HS – hot spot), the response format, and the scoring procedure.

Table 1. Number and Types of Items on the ACCESS 601 Listening Test

Grade-Level Cluster	Tier Pool	Number of Items	Targeted PL range	Item Types and Percentages*			Response Formats	Scoring Procedures
				MC	DD	HS		
1	Entry	6	PL1–PL4	83%	0%	17%	Dichotomous selected response	Machine scored
1	A	12	PL1–PL3	100%	0%	0%		
1	B	18	PL2–PL4	77%	6%	17%		
1	C	18	PL3–PL5	100%	0%	0%		
2–3	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
2–3	A	12	PL1–PL3	100%	0%	0%		
2–3	B	18	PL2–PL4	78%	0%	22%		
2–3	C	18	PL3–PL5	89%	0%	11%		
4–5	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
4–5	A	12	PL1–PL3	100%	0%	0%		
4–5	B	18	PL2–PL4	83%	0%	17%		
4–5	C	18	PL3–PL5	100%	0%	0%		
6–8	Entry	6	PL1–PL4	83%	0%	17%	Dichotomous selected response	Machine scored
6–8	A	12	PL1–PL3	83%	0%	17%		
6–8	B	18	PL2–PL4	55%	17%	28%		
6–8	C	18	PL3–PL5	94%	0%	6%		
9–12	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
9–12	A	12	PL1–PL3	92%	0%	8%		
9–12	B	18	PL2–PL4	77%	6%	17%		

Grade-Level Cluster	Tier Pool	Number of Items	Targeted PL range	Item Types and Percentages*			Response Formats	Scoring Procedures
				MC	DD	HS		
9–12	C	18	PL3–PL5	100%	0%	0%		

*Item types are MC – multiple choice; DD – drag-and-drop; HS – hot spot.

The Listening test uses a multistage adaptive design, as illustrated in Figure 16. All students begin the Listening test with two entry folders (with three items each) at Stage 1 and Stage 2, both targeting Social and Instructional Language (see Section 1.2 for the WIDA ELD Standards). At that point, the test engine estimates the student's ability based on the student's performance on those six items, and the engine then uses that ability estimate to determine which of the three leveled Language of Language Arts folders in Stage 3 is administered next. Students whose ability estimate predicts a PL score of 5.0 or higher are routed into the folder at the highest level (C in Figure 16); students whose ability estimate predicts a PL score of 2.5 or lower are routed into the folder at the lowest level (A in Figure 16); all others are routed into the B folder. Throughout the test, the test engine re-estimates a student's underlying measure of ability with the completion of each folder, and the engine then uses that information to choose the level of the next folder to be administered, following the decision rules above. Thus, each student will trace a tailor-made path through the test according to ability level, but the order of the stages is invariant across students. In total, there are eight possible stages, but a student whose ability estimate falls below PL 2.5 after the sixth stage ends the test at that point. This shortening of the test for students at the lower proficiency levels allows them to demonstrate what they know without subjecting them to additional content, when their ability is not near the cut point where the EL reclassification decision is made. The intent of this design is to ensure coverage of the Standards while delivering a test that closely matches the student's PL, thus minimizing measurement error. Although timing guidance is included in the Test Administrator Manual (WIDA Consortium, 2021a), the Listening test is untimed.

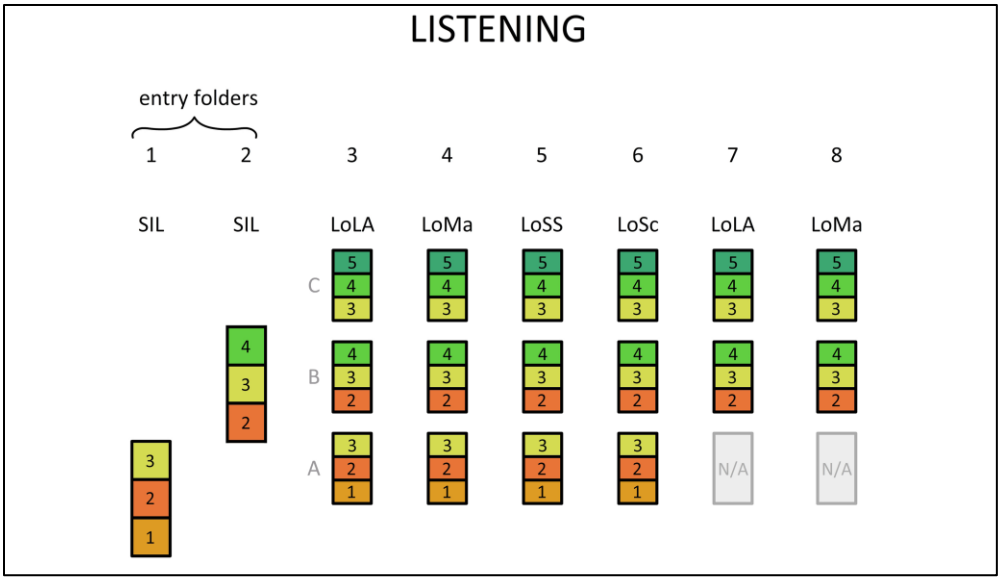


Figure 16. Format of the Listening test

2.2.2 Reading

For the ACCESS Reading test, Table 2 shows, for each grade-level cluster and tier pool, the number of items, the targeted range of WIDA proficiency levels, the proportion of items by item type (MC – multiple choice; DD – drag-and-drop; HS – hot spot), the response format, and the scoring procedure.

Table 2. Number and Types of Items on the ACCESS 601 Reading Test

Grade-Level Cluster	Tier Pool	Number of Items	Targeted PL range	Item Types and Percentages*			Response Formats	Scoring Procedures
				MC	DD	HS		
1	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
1	A	18	PL1–PL3	100%	0%	0%		
1	B	24	PL2–PL4	96%	0%	4%		
1	C	24	PL3–PL5	100%	0%	0%		
2–3	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
2–3	A	18	PL1–PL3	100%	0%	0%		
2–3	B	24	PL2–PL4	100%	0%	0%		
2–3	C	24	PL3–PL5	100%	0%	0%		
4–5	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
4–5	A	18	PL1–PL3	94%	0%	6%		
4–5	B	24	PL2–PL4	96%	4%	0%		
4–5	C	24	PL3–PL5	100%	0%	0%		
6–8	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
6–8	A	18	PL1–PL3	100%	0%	0%		
6–8	B	24	PL2–PL4	100%	0%	0%		
6–8	C	24	PL3–PL5	100%	0%	0%		
9–12	Entry	6	PL1–PL4	100%	0%	0%	Dichotomous selected response	Machine scored
9–12	A	18	PL1–PL3	100%	0%	0%		
9–12	B	24	PL2–PL4	100%	0%	0%		

Grade-Level Cluster	Tier Pool	Number of Items	Targeted PL range	Item Types and Percentages*			Response Formats	Scoring Procedures
				MC	DD	HS		
9–12	C	24	PL3–PL5	100%	0%	0%		

*Item types are MC – multiple choice; DD – drag-and-drop; HS – hot spot.

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

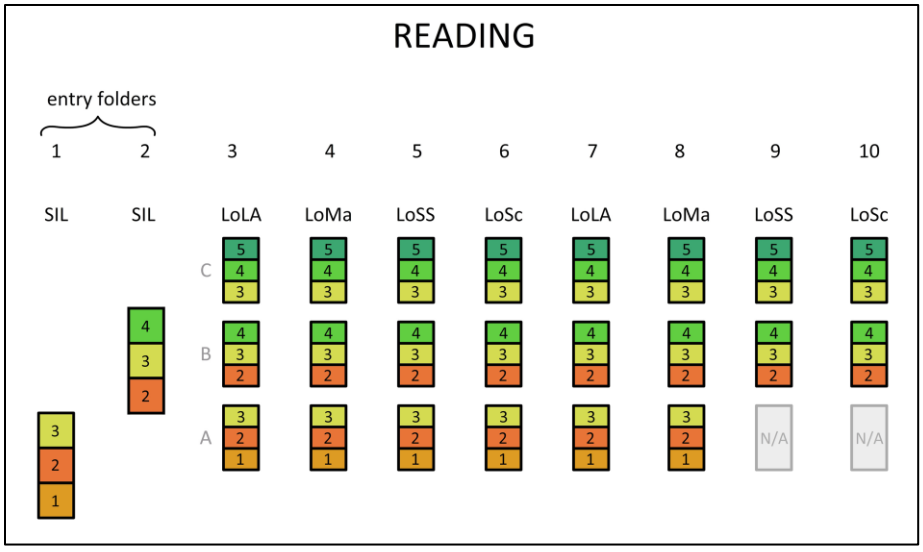


Figure 17. Format of the Reading test

2.2.3 Writing

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

Table 3. Number and Types of Tasks on the Writing Test

Grade-Level Cluster	Tier	Number of Tasks	Targeted PL Range	Task Type	Response Formats	Scoring Procedures
1	A	2	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
1	B/C	2	PL2–PL5			
2–3	A	2	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in test booklet	Human scored: centrally scored by DRC
2–3	B/C	2	PL2–PL5			
4–5	A	2	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in response booklet or keyboarded in test platform	Human scored: centrally scored by DRC
4–5	B/C	2	PL2–PL5			
6–8	A	2	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in response booklet or keyboarded in test platform	Human scored: centrally scored by DRC
6–8	B/C	2	PL2–PL5			
9–12	A	2	PL1–PL3			

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

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

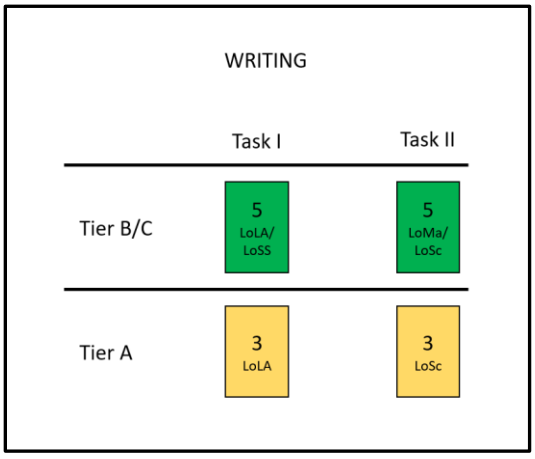


Figure 18. Format of the Writing test

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

The design of the embedded Writing field test for Series 601 is described in greater detail in Section 2.3.2.3 below.

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

Although timing guidance is included in the Test Administrator Manual, the Writing test is untimed.

2.2.4 Speaking

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

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

Grade-Level Cluster	Tier	Number of Tasks	Targeted PL range	Task Type	Response Formats	Scoring Procedures
1	Pre-A	3	PL1	Speaking constructed response	Polytomous constructed response	Human scored; centrally scored by DRC
1	A	6	PL1–PL3			
1	B/C	6	PL3–PL5			
2–3	Pre-A	3	PL1	Speaking constructed response	Polytomous constructed response	Human scored; centrally scored by DRC
2–3	A	6	PL1–PL3			
2–3	B/C	6	PL3–PL5			
4–5	Pre-A	3	PL1	Speaking constructed response	Polytomous constructed response	Human scored; centrally scored by DRC
4–5	A	6	PL1–PL3			
4–5	B/C	6	PL3–PL5			
6–8	Pre-A	3	PL1	Speaking constructed response	Polytomous constructed response	Human scored; centrally scored by DRC
6–8	A	6	PL1–PL3			
6–8	B/C	6	PL3–PL5			
9–12	Pre-A	3	PL1	Speaking constructed response	Polytomous constructed response	Human scored; centrally scored by DRC
9–12	A	6	PL1–PL3			
9–12	B/C	6	PL3–PL5			

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

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

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

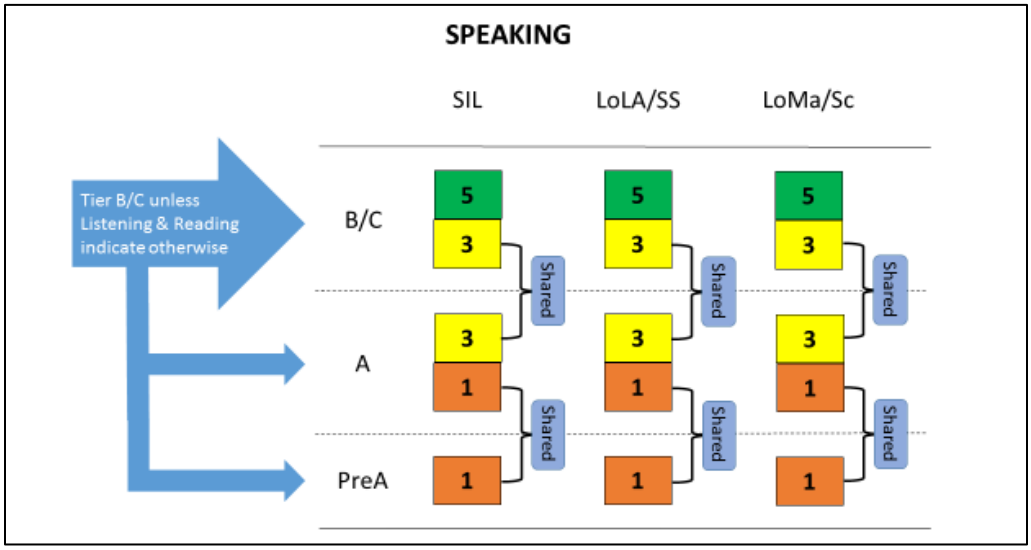


Figure 19. Format of the Speaking test

As with the Writing test, placement of students into the three tiers on the Speaking test depends on their performance on the Listening and Reading tests. Unlike Writing, the Speaking test has one additional tier, Tier Pre-A. Students are placed into Tier Pre-A when their scores on both the Listening and Reading tests are below PL 2.0. The Speaking Pre-A tier is designed to meet the needs of students in the very early stages of English language development. As noted previously, these tasks are targeted to the P1 level. DRC raters score students' responses to these tasks using a modified version of the full Speaking rating scale (see Section 3.2.4).

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

Although timing guidance is included in the Test Administrator Manual, the Speaking test is untimed.

2.3 Test Construction

2.3.1 Item and Task Development

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

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

- The CAL TD team analyzes prior items and tasks that could not be used operationally due to fit issues, or in cases where the item or task fit was acceptable, an item or task difficulty measure that was outside of the range for the intended slot on the test. The purpose of this analysis is to determine if the poor performance is due to item/task

mechanics (e.g., an issue with the wording of the passage or stem, a distractor that is too attractive) or if there is a deeper item/task-specification issue that cannot be resolved (e.g., the specification is difficult to operationalize successfully). In the latter case, the CAL TD team can update the specification (usually focused on updating the MPIs) or completely replace it, depending on the specific situation.

- The CAL TD team also updates or replaces item/task specifications as content standards change. As noted previously, the ACCESS item/task specifications include explicit connections to the content standards. If an update to the relevant content standard makes an ACCESS item/task specification obsolete, the CAL TD team revises or replaces the specification.

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

The CAL TD team and WIDA TD staff are responsible for recruiting classroom English as a second language (ESL) and content teachers with experience teaching the academic content associated with one or more of the WIDA ELD Standards (including educators with experience working with English learners with disabilities), and the CAL TD team provides these educators with key parts of the item/task specification document (i.e., the topic, the MPIs, and guidelines for selecting a good theme). Then, the CAL TD team asks educators to propose themes related to the topic, along with possible directions for each item or task, which are grade-level appropriate. After the theme generation process is complete, the CAL TD team reviews the list of themes to identify those that will become the focus of item/task writing. This determination is based on several factors, including operationalizability on a large-scale assessment (since many ideas from educators are well suited for the classroom but do not clearly translate to the assessment context), themes currently in use on the assessment, and bias and sensitivity considerations.

The team then assigns themes to professional item/task writers to develop the initial item/task content. The team recruits individuals with prior experience developing ESL or English language

arts items/tasks, preferably in the context of large-scale, standardized assessments, but individuals with other experience (such as experience writing items/tasks for language tests in languages other than English, and experience with English placement tests for the college/university setting) are also considered. All item/task writers, both new item/task writers and those returning from the previous test development cycle, participate in an introductory training, and the team provides them with extensive documentation regarding writing items/tasks for ACCESS, including an Item Writing Handbook and ancillary documents (i.e., checklists, item/task specifications, templates) to complete their assignments. One or more CAL Language Testing Specialists work with each item/task writer, providing feedback on the item content.

After item/task writing is complete, CAL Language Testing Specialists and Test Development Managers review the folders, using a standard checklist, to determine which folders will undergo further development and which will be retired. Folders then go to their first external review, the Standards Expert review.

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

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

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

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

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

The Bias and Sensitivity Review Panel ensures that test items and tasks are free of material that (1) might favor any subgroup of students over another on the basis on gender, race/ethnicity, home language, religion, culture, region, or socioeconomic status, and (2) might be upsetting to students. Educators conduct Bias and Sensitivity Reviews by grade groupings (e.g., G1-3, G4-5, G6-8, and G9-12). The educators who are recruited to review a particular grade cluster's content (5 or 6 reviewers per grade grouping) are educators or school administrators who have experience teaching English language learners and are either currently teaching students who are in that grade cluster or have extensive prior experience teaching students who are in that grade cluster. WIDA TD staff employ additional criteria to ensure that a variety of perspectives are represented on each panel. These criteria include recruiting at least one educator with experience in Special Education to serve on each panel. Additionally, during the recruitment process, WIDA TD staff seek to ensure diversity and balance across (1) consortium states, (2) school locale (rural/suburban/urban), and (3) years of teaching experience. The CAL TD team and WIDA TD staff conduct training for all new and returning reviewers before any items or tasks are reviewed. The panel members then individually review each item and task, followed by a collective discussion of each item and task to determine if any bias or sensitive topics are detected in the items/tasks, and if so, what the CAL TD team can do to remediate the issues. The CAL TD team and WIDA TD staff facilitate the reviews and take extensive notes to capture all feedback during the reviews. WIDA TD staff also conduct a separate, asynchronous review around the time of the Content Review and Bias and Sensitivity Review, using the same materials that the educators review, and provide written feedback on the materials.

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

Tasks in the Writing domain undergo one additional step: a small-scale tryout with educators and students. Given the changes to the Writing test over the past few years, including a change from three to two operational tasks, along with changes to task specifications to better align the Writing tasks with classroom practice, these tryouts allow the CAL TD team to evaluate whether

each Writing task will effectively elicit language at its targeted WIDA proficiency level. For the Writing tryouts, the CAL TD team and WIDA TD staff jointly recruit educators with appropriate numbers of students at the targeted proficiency levels (approximately 15 students per task) to participate. The CAL Test Development Manager for Writing prepares a recruitment flyer, which the CAL TD team and WIDA TD staff circulate to educators. The CAL TD team circulates the flyer to educators who have previously participated in the tryouts, and WIDA TD staff circulate the flyer through WIDA's regular SEA/LEA communications emails. Due to the small-scale nature of the tryouts, the recruitment is ultimately a convenience sample, although CAL and WIDA strive to obtain a sample with geographic diversity (i.e., educators distributed throughout the consortium, with a mix of urban, suburban, and rural representation). In addition, the tasks target different proficiency levels/tiers at the different grade level clusters, so recruiting educators with students at these targeted proficiency levels and grade level clusters is another requirement. Finally, we do not recruit educators who have already reviewed the tasks in development during Bias and Sensitivity and Content review to participate in tryouts. If the CAL TD team determines that the first round of recruitment has failed to find educators with students at the appropriate proficiency levels for all grade clusters and tiers, the CAL Test Development Manager for Writing identifies the grade-level clusters and proficiency levels/tiers with gaps and provides this information to the WIDA TD staff, who can then do targeted recruitment based on existing databases of educators who have indicated willingness to participate in test development activities.

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

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

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

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

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

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

In recent years, WIDA has placed additional focus on ensuring that the items and tasks, and especially the graphics, are amenable to accommodations by involving WIDA's Accessibility and Accommodations Team directly in the item/task review process. WIDA's Accessibility and Accommodations Team helped CAL's TD team develop principles for graphics development and for eliminating language that is biased towards students with sight, and WIDA's Accessibility and Accommodations Team also reviews the items and tasks during development to help CAL identify areas that still need to be addressed.

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

2.3.2 Field Testing

2.3.2.1 Listening

DRC field tested the Listening items developed for Series 601 as embedded folders during the operational administration of Series 503. The embedded field test folders contained items that featured innovative formats, including hot spot items (i.e., the student clicks on an area of the screen to respond) and drag-and-drop items (i.e., the student drags an image/text to a specified screen area to respond).

For Series 601, DRC field tested a total of 120 Listening items (40 folders), across all five grade-level clusters, as indicated in Table 5.

Table 5. Number of Field Test Folders and Items for the Series 601 Listening Test

Grade-Level Cluster	Tier Pool	Number of Folders to Refresh	Number of Overage Folders	Total Number of Field Test Folders	Total Number of Field Test Items	Standards Addressed in FT
1	Entry	1	0	1	3	SI
1	A	1	1	2	6	MA
1	B	2	1	3	9	LA, SC
1	C	1	1	2	6	MA
2–3	Entry	1	0	1	3	SI
2–3	A	2	2	4	12	MA, SC
2–3	B	1	1	2	6	LA
2–3	C	1	1	2	6	SS
4–5	Entry	1	0	1	3	SI

Grade-Level Cluster	Tier Pool	Number of Folders to Refresh	Number of Overage Folders	Total Number of Field Test Folders	Total Number of Field Test Items	Standards Addressed in FT
4–5	A	2	1	3	9	MA, SC
4–5	B	2	0	2	6	SS, SC
4–5	C	0	0	0	0	
6–8	Entry	1	0	1	3	SI
6–8	A	1	1	2	6	MA
6–8	B	1	0	1	3	MA
6–8	C	2	2	4	12	MA
9–12	Entry	1	0	1	3	SI
9–12	A	0	0	0	0	
9–12	B	2	2	4	12	LA, SS
9–12	C	2	2	4	12	LA, SS
Total		25	15	40	120	

Each student received one Listening field test folder embedded in the operational test. Field test folders are targeted to refresh a specific operational folder on the test, and field test folder specifications include the stage, WIDA ELD Standard, and tier pool target (i.e., Entry, A, B, or C) of the folder. Students received the embedded field test folder at the stage targeted for refreshment, with administration randomized so that half of the students saw the field test folder before the corresponding operational folder, and half saw the operational folder before the field test folder. Field test folders were administered to those students who were routed to take the operational folder that was either at the same tier or adjacent to the tier that the field test folder targeted. When DRC drew the field test samples, 50% of the sample were students who were routed to the tier that the field test folder targeted, and the other 50% were students who were

routed to adjacent tiers. (If there were adjacent tiers both above and below the field test target, then 25% of the sample were students routed to each of those tiers.) In cases where the field test folder was to be placed in one of the entry stages, students receiving that field test folder took it directly after the pair of operational entry folders. CAL set the field test sample targets for the Listening test at a minimum of 3,000 responses per folder.

Because CAL's psychometrics team used the Listening field test data in the pre-equating analysis, their sample size requirement of 3,000 was much higher than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, to ensure that the pre-equated parameter estimates would be stable. Linacre (1994), citing Wright and Douglas's (1977) formulation, explained how to determine the minimum sample required for calibrating dichotomous items to achieve various levels of estimation precision and confidence intervals. With a sample size of 3,000, one can be 95% confident that no item parameter will be more than ±0.1 logit away from its true value. The sample sizes for all field test folders exceeded the minimum requirement of 3,000, except for one Listening grade Cluster 4–5 Tier A folder, which had a sample size of 2,800, due to the fact that the population of grade 4-5 tier A students comprise a smaller portion of the population than students in other grade level clusters and tiers.

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

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

Table 6. CAL's Post–Field Test Review Classification System for Series 601

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

2.3.2.2 Reading

DRC field tested the Reading items developed for Series 601 as embedded items during the operational administration of Series 503. The embedded field test folders contained items that featured innovative formats, including hot spot items (i.e., the student clicks on an area of the screen to respond) but no drag-and-drop items.

For Series 601, DRC field tested a total of 201 Reading items (67 folders), across all five grade-level clusters, as indicated in Table 7.

Table 7. Number of Field Test Folders and Items for the Series 601 Reading Field Test

Grade-Level Cluster	Tier Pool	Number of Folders to Refresh	Number of Overage Folders	Total Number of Field Test Folders	Total Number of Field Test Items	Standards Addressed in FT
1	Entry	1	0	1	3	SI
1	A	2	2	4	12	MA,SC
1	B	2	1	3	9	LA, SS

Grade-Level Cluster	Tier Pool	Number of Folders to Refresh	Number of Overage Folders	Total Number of Field Test Folders	Total Number of Field Test Items	Standards Addressed in FT
1	C	1	1	2	6	SS
2–3	Entry	1	0	1	3	SI
2–3	A	0	0	0	0	
2–3	B	3	2	5	15	LA, SC, SS
2–3	C	3	3	6	18	LA, MA,
4–5	Entry	1	0	1	3	SI
4–5	A	2	0	2	6	SC, LA,
4–5	B	3	1	4	12	SC, MA
4–5	C	5	4	9	27	MA, SS, LA, SC
6–8	Entry	1	0	1	3	SI
6–8	A	1	0	1	3	MA
6–8	B	4	1	5	15	LA, SS, SC
6–8	C	4	4	8	24	SC, MA, SS
9–12	Entry	1	0	1	3	SI
9–12	A	1	0	1	3	SS
9–12	B	4	3	7	21	MA, SS, SC
9–12	C	3	2	5	15	MA, LA, SC
Total		43	24	67	201	

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

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

2.3.2.3 Writing

DRC administered the Series 601 Writing tasks in an embedded field-test model. For Series 601, a total of 17 Writing tasks were field tested, as indicated in Table 8.

Table 8. Number of Field Test Tasks for Series 601 Writing

Grade-Level Cluster	Tier	Number of Folders to Refresh	Number of Folders Field Tested	Standards Addressed in FT
1	A	1	2	SC
1	BC	1	2	MA/SC
2–3	A	1	2	SC
2–3	BC	1	2	MA/SC
4–5	A	1	1	SC
4–5	BC	1	2	MA/SC
6–8	A	1	1	SC
6–8	BC	1	2	MA/SC
9–12	A	1	1	SC
9–12	BC	1	2	MA/SC
Total		10	17	

All students received a field test folder that was appended to their operational assessment. Students received a field test folder in the tier that corresponded to their operational tier. CAL targeted a sample of 500 students per task. This was much higher than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, making it likely that, for each of the nine scale categories, there would be at least 10 students whose responses to the task would warrant

receiving scores in that category, as Linacre (2002a) recommended for polytomously scored tasks. If raters assign fewer than 10 scores in each scale category, then the category statistics for that category tend to be unstable. Historically, the distribution of scores that raters assign to students' responses to the Writing tasks tends to be highly concentrated in the middle of the score distribution (i.e., exhibit a central tendency effect), with raters assigning relatively fewer scores in the categories at the high end of the score scale. Therefore, CAL targeted a sample size of 500 to ensure that there would be students for analysis whose responses to the task would warrant receiving scores at the high end of the 9-category score scale. Use of this larger sample size also provided examples of students' responses that received scores in the higher scale categories that trainers could use as anchor papers for rater training.

DRC administered the field test under standard testing conditions. The field test used the online interface with keyboarded responses for Grades 4–12 and paper booklets with handwritten responses for Grades 1–3. For the Writing field test, DRC raters scored the students' responses to the field test tasks. DRC performed a 20% read-behind as a quality control measure, with the first score assigned as the score of record.[4]

2.3.2.4 Speaking

All Tier A and B/C students received a Speaking field test folder that was appended to their operational Speaking assessment. Tier Pre-A was not included in the field test. DRC field tested a total of 30 folders (15 panels) for Series 601, with a target sample size of 500 students per folder. This is much higher than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, making it likely that, for each of the four scale categories, there would be at least 10 students whose responses to the task would warrant receiving scores in that category, as Linacre (2002a) recommended for polytomously scored tasks. Historically, the distribution of scores that raters assign to students' responses to the Speaking tasks tends to be highly concentrated in the middle of the score distribution (i.e., exhibit a central tendency effect), with

[4] The purpose of the 20% read-behind is to monitor rater performance on a daily basis. (See Section 3.2.2 below). If the read-behinds detect that one rater is consistently scoring inaccurately, DRC can rescore all of the students' responses to tasks scored by that rater, and the rater can be retrained or terminated. Raters go through significant training and qualification prior to live scoring, and they are monitored daily through validity and recalibration tasks, so a scenario where a rater is consistently anomalous in his or her ratings would be uncommon, and it would be detected and corrected immediately.

raters assigning relatively fewer scores in the categories at the high end of the score scale. Therefore, CAL targeted a sample size of 500 to ensure that there would be students for analysis whose responses to the task would warrant receiving scores at the high end of the 4-category score scale.[5] Use of this larger sample size also provided examples of students' responses that received scores in the higher scale categories that trainers could then use as anchor papers for rater training.

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

Students received a Speaking field test folder in the tier that corresponded to their operational tier. For Series 601, CAL field tested a total of 30 Speaking folders, as indicated in Table 9.

Table 9. Number of Field Test Tasks for Series 601 Speaking

Grade-Level Cluster	Tier	Number of Folders to Refresh	Number of Folders Field Tested	Standards Addressed in FT
1	A	2	3	SIL, LA/SS
1	BC	2	3	SIL, LA/SS
2–3	A	2	3	SIL, LA/SS
2–3	BC	2	3	SIL, LA/SS
4–5	A	2	3	SIL, LA/SS
4–5	BC	2	3	SIL, LA/SS
6–8	A	2	3	SIL, LA/SS
6–8	BC	2	3	SIL, LA/SS
9–12	A	1	2	LA/SS
9–12	BC	1	2	LA/SS
Total		20	28	

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

2.3.3 Item/Task Review and Selection

After the analysis of field test data, a panel consisting of members of the WIDA TD and psychometrics staff, the CAL TD Team, and the CAL psychometrics team conducted an item/task selection meeting to determine which of the field-tested folders would be placed on the Series 601 operational assessment. Results from qualitative and quantitative analyses guided the selection of operational items and tasks.

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

Items were coded by color according to the following criteria:

- If an item showed C-level or CC-level differential item functioning (DIF), it was automatically coded yellow. Any items that showed this level of DIF were subject to an extra round of review (to determine if anything in the item could be detected that clearly indicates bias) prior to item selection (see Part 2, Section 2.2 for further detail). The CAL psychometrics team provided the Item Selection Panel with the report of the DIF review.
- Items were coded as green if they had infit and outfit values ≤1.20. As outfit and infit values are sensitive to students' unexpected responses to items that are very easy for them, any item with a *p* value >0.85 was automatically coded as green, even if it had fit values outside of these thresholds.
- Items with infit and outfit values >1.20 and <1.50 were coded as yellow. As outfit values are also sensitive to students' unexpected responses to items that are very hard for them, items with *p* values close to chance (0.40 for a three-response item and 0.35 for a four-response item) were coded as yellow if outfit was >1.20 and <1.75.
- Items that did not meet these criteria were coded as red.

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

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

Table 10 and

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

Table 10. Number of New and Continuing Items on ACCESS Online, Series 601 Listening Test, by Grade-Level Cluster

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

Table 11. Number of New and Continuing Items on ACCESS Online, Series 601 Reading Test, by Grade-

Level Cluster

Grade-Level Cluster	Number of New Items	Number of Continuing Items	Total Number of Items
1	18	54	72
2–3	21	51	72
4–5	9	63	72
6–8	15	57	72
9–12	15	57	72

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

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

were generally rejected if the task difficulty measures and the raw score distributions of the field-test tasks varied too much from those of the operational tasks.

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

Table 12. Number of New and Continuing Tasks on ACCESS Online Series 601 Writing Test, by Grade-Level Cluster

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

Table 13. Number of New and Continuing Tasks on ACCESS Online Series 601 Speaking Test, by Grade-Level Cluster

Grade-Level Cluster	Tier	Number of New Tasks	Number of Continuing Tasks	Total Number of Tasks
1	Pre-A	1	2	3

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

3. Test Administration

3.1 Test Delivery

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

3.1.1 Listening and Reading

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

3.1.2 Writing

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

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

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

3.1.3 Speaking

Speaking tasks are delivered online. Students listen to prompts via headsets that are equipped with microphones to capture their responses. The student receives extensive support via

illustrations and multimodal (text and audio) input designed to provide sufficient context for the response, as well as a model student response that provides guidance on the level of linguistic complexity required to respond adequately (see Section 2.2.4).

3.2 Operational Administration

Before, during, and after a state's testing window, there are various roles that educators hold to ensure all tasks are carried out for successful test administration. These roles include Test Coordinators at the district and school level, Test Administrators, and, for online administration, Technology Coordinators. The Test Administrator administers and monitors the test and is responsible for managing student data prior to, during, and after testing. The *Test Administrator Manual* and the *District and School Test Coordinator Manual* contain more information related to responsibilities and required training for the various roles. These manuals can be found on the WIDA Secure Portal (https://portal.wida.us/).

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

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

3.2.1 Administering the Test Practice

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

captioning displayed directly on the screen, allowing the student to be able to read along as the script is read aloud.

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

3.2.2 Listening Test Administration

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

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

3.2.3 Reading Test Administration

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

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

3.2.4 Writing Test Administration

All students in Grades 1–3 complete the ACCESS for ELLs Writing test on paper. The test is group administered. For Grades 6–12, all students view the Writing prompts on the desktop, laptop, or tablet. The default response mode is keyboarding. For Grades 4–5, all students also

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

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

3.2.4.1 Writing Test Tiers

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

3.2.5 Speaking Test Administration

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

Recording response time on every task on the Speaking test has a preset time limit, which varies depending on the grade-level cluster, tier, and task level. Students learn about the time limits in the test directions and practice. Students see a circle change color and then disappear as the time to respond elapses. While there is a limit to how long students can take to record their response, students can navigate the directions, practice, and test items at their own pace. Students click the

Next button when they are ready to move on from a screen, without time limits. The test does not advance automatically.

3.2.5.1 Speaking Test Tiers

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

3.2.5.2 Group vs. Individual Delivery

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

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

3.2.6 Test Security

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

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

3.3 Fairness and Accessibility

The WIDA Accessibility and Accommodations Framework provides support for all ELLs, as well as targeted accommodations for students with individualized education plans (IEPs) or 504 plans. These supports are intended to increase the accessibility for the assessments for all ELLs. (Please see Accessibility and Accommodations Supplement for detailed information: https://wida.wisc.edu/resources/accessibility-and-accommodations-supplement). Fairness and accessibility are considered throughout the assessment process (i.e., test design, test development, item selection, forms creation, and test administration). For details, please refer to the universal design principles throughout test and item design to the *WIDA consortium English Language Proficiency Assessment for grades 1-12 Test and Item Design Plan ACCESS for ELLs Online Annual Summative Assessment and WIDA Screener Online.*

3.3.1 Support Provided to All ELLs

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

Administrative considerations include adaptive and specialized equipment or furniture, alternative microphone, familiar Test Administrator, frequent or additional supervised breaks, individual or small group setting, monitoring of the placement of responses in the test booklet or on screen, participation in different testing formats (Paper vs Online), reading aloud to self, specific seating, short segments, verbal praise or tangible reinforcement for on-task or appropriate behavior, and verbal redirection of students' attention to the test (in English or native language).

Universal tools are available to all students taking ACCESS for ELLs to address their accessibility needs. These may either be embedded in the online test or provided by Test Administrators during testing. The universal tools provided on ACCESS for ELLs Online are described in Section 3.3.2 below. The Test Demo videos available on the WIDA website

(https://wida.wisc.edu/assess/access/preparing-students/practice) instruct students how to use the universal tools. During testing, students choose whether to use the tools or not, but they are available to all students throughout testing.

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

WIDA also offers Braille Test for ELLs and Large Print Test. 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 ELLs who are blind. The Large Print Test is used for students with visual impairments. The font size on the large print paper test is increased to 18 point. For the online test, the magnification/zoom tool increases the on-screen font size up to 1.5× or 2×, depending on the size of the computer monitor.

Universal tools are also available to all ELLs taking ACCESS for ELLs. Examples of universal tools include highlighter, line guide, magnification, and color overlay. All universal tools are available to all ELLs during testing; specific designation is not required prior to testing to make them available to the student during testing. The Test Demo videos available on the WIDA website (https://wida.wisc.edu/assess/access/preparing-students/practice) instruct students how to use the universal tools. During testing, students choose whether to use the tools or not, but they are available to all students throughout testing. Features available during online-based test administration include the following:

- Audio amplification device (provided by student)
- Highlight tool

- Line guide
- Zoom tool (magnifier)
- Sticky notes—which allow students to take notes to prepare responses to Writing items. This tool is only available in the Writing domain.
- Color overlay—which allows students to change the background color that appears behind text, graphics, and response areas. Five colors are available: pink, yellow, blue, green, and orange.
- Color contrast—which allows students to select from a variety of background/text color combinations
- Keyboard shortcuts/equivalents—which are alternatives to using a mouse (for navigating through the test and using online test tools)
- Scratch/blank paper (to be submitted with the test or disposed of according to state policy)

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

Schools and districts should consider how accessibility features and allowable test administration procedures can support accessibility to the test for *all* ELLs. 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 ELLs and students with disabilities; and (4) the expertise of the CAL test developers.

WIDA offers *Alternate ACCESS for ELLs.* This test is intended only for those ELLs who have cognitive disabilities that are so significant as to prevent meaningful participation in ACCESS

testing, even with accommodations. The results of the Alternate ACCESS for ELLs operational administration appear in a separate technical report.

4. Scoring

4.1 Multiple Choice Scoring: Listening and Reading

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

4.2 Scoring Performance-Based Tasks: Writing and Speaking

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

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

Preparing Rater Training Materials for Speaking and Writing tasks

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

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

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

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

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

Finally, the TD Manager reviews the score explanations to check that they meet the required criteria. The TD Manager also selects 20 responses from the initial review to be used as training samples, and reviews and revises any accompanying score notes as necessary. The 20 training

samples are selected so that the full range of observed score points are included in the set, and so that the most commonly observed score points for the grade level cluster and tier are well-represented. The TD Manager also reviews all notes from the anchor and training sample selection process and, when necessary, compiles any task-specific scoring guidance to be used by raters.

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

Following field test scoring and operational item selection, CAL adds to the rater training materials for each task that was selected for the operational test. This primarily consists of selecting and annotating additional training samples, so that a minimum of 30 samples are provided for operational rater training. In some cases, additional anchor responses are also added to the anchor set, when an appropriate anchor response for the highest observed score point was not found while preparing for field test scoring but could be identified once a larger pool of scored responses was available.

Rater Training and Qualifying

- DRC assigned each rater a unique ID number and password.
- The scoring director conducted a team leader training session before training the raters. This session followed the same procedures as rater training but was more rigorous and in-depth due to the extra responsibilities required of team leaders. During team leader training, all WIDA materials were reviewed and discussed. To facilitate scoring consistency, it was imperative that all team leaders imparted the same rationale for each response. Once the team leaders were qualified, leadership responsibilities were reviewed, and team assignments were given.
- Rater training began with the scoring director going through the ACCESS for ELLs PowerPoint presentation provided by CAL. The PowerPoint gave scorers a good overview of ACCESS for ELLs and the WIDA scoring process.
- Rater training continued with the scoring director providing an intensive review of the ACCESS for ELLs Scoring Scale, the model student response for Speaking items, and task-specific anchor sets created by CAL. The anchor set contained a collection of student responses that were used to exemplify each possible score point. Each response

included a scoring annotation that explained the scoring rationale. Scorers used the ACCESS for ELLs Scoring Scale, the model student response for Speaking, and the anchor sets as primary references during scoring.

- Next, raters practiced by independently scoring responses in training sets. Training sets were created by DRC scoring directors from responses approved by WIDA and CAL. The responses were selected to show raters the range of each score point (e.g., high, mid, and low 2s). This process helped raters recognize the various ways that a student could respond in order to earn each score point outlined and defined in the scoring guidelines. After each training set was taken, the scoring director led a thorough discussion of the responses.
- Once the scoring scale, anchor sets, and training sets were thoroughly discussed, each rater was required to demonstrate understanding of the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the true scores) on at least one of the qualifying sets. Raters who failed to achieve at least 70 percent exact agreement on the first qualifying set were given additional training, either individually or in a small group setting. Raters who did not perform at the required level of agreement by the end of the qualifying process were not allowed to score any student responses. These individuals were removed from the pool of potential raters in DRC's imaging system and released from the project. Qualifying sets were created by DRC scoring directors from responses approved by WIDA and CAL.
- Throughout training, the scoring director provided detailed directions for use of DRC's computerized scoring system and remote communication tools for raters.
- Once raters were trained, qualified, and began live scoring, DRC used recalibration sets and validity responses to keep the raters calibrated on the tasks they were scoring. Recalibration sets were pre-scored sets of responses that were approved by WIDA and CAL and were used to help refocus raters on WIDA scoring guidelines. Validity responses were also approved by WIDA and CAL and were responses that were pre-scored and used to ensure raters were adhering to WIDA scoring criteria. Recalibration and validity are explained in greater detail below.

Calculating Score Agreement for Score Monitoring

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

Routing Responses to Ensure "Blind" Second Scores

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

Monitoring Scoring (Quality Control)

- Rater accuracy was monitored throughout the scoring session by means of daily and on-demand reports. These reports ensured that an acceptable level of scoring accuracy was maintained throughout the project. Interrater reliability was tracked and monitored with multiple quality control reports. These reports and other quality control documents were generated at the scoring centers, where they were reviewed by the scoring directors, team leaders, and project managers. DRC provided WIDA with access to these reports on a regular basis throughout the scoring process to provide assurance that the quality control metrics met or exceeded expectations. If a scorer did not meet scoring expectations, a portion of, or all, their scores could be dropped if the scores had not been reported.
- During the handscoring process, the scoring directors communicated regularly with their team leaders to review the statistics generated from the previous day's work, including interrater reliability, score point distributions, and validity reports.
- Throughout handscoring, team leaders conducted routine read- and listen-behinds to observe, in real time, raters' performance. Team leaders utilized live, scored responses to provide ongoing feedback and, if necessary, retraining for raters.
- The DRC system generated interrater reliability reports daily to monitor how often each rater's scores matched other raters' scores, and scoring leaders continually monitored individual rater statistics, comparing them to the group average. If the agreement rate for a rater fell below 70%, supervisors increased monitoring and retraining activities with the rater. If the rater failed to demonstrate improved reliability, DRC released the rater from scoring responses to that task.
- Since the interrater agreement rates were all at or above 70%, the target that WIDA stipulated, the focus turned to raters with lower-than-average agreement rates—even if their agreement rate was at or above 70%. Even when all agreement rates were at or above 70%, scoring supervisors continued to seek opportunities to increase reliability by providing ongoing feedback and retraining raters based on the specific performance of each rater, as evidenced by the quality control reports and observations made when reviewing scores that a rater assigned.
- DRC can retrieve students' responses on demand (e.g., specific grade-level clusters, specific students) should the need arise during or after the scoring process.
- If needed, DRC can re-score a student's response to a task based on task- or response-level information, such as task number, date, score assigned, or rater ID.

- For both Speaking and Writing, DRC used both recalibration sets and validity responses to monitor handscoring quality control. DRC, CAL, and WIDA collaborated to develop these recalibration sets and validity responses. CAL developed an initial pool of responses for use as recalibration and validity checks by selecting responses from a previous administration of the tasks (e.g., a field test). WIDA staff reviewed and approved this pool of responses and their scores. DRC supervisors supplemented this pool of responses as needed by selecting additional responses, which CAL and WIDA approved before use. For each of the first 5 days that raters scored student responses to a task, they scored one recalibration set of five responses. The recalibration sets did not differ from rater to rater. For example, DRC identified a recalibration set to use for the first day that a rater scored students' responses to a specific task; every rater who was working on that task took this same recalibration set on the first day that they worked on that task. After the raters assigned scores to the recalibration set, the scoring director or team leader reviewed the set using descriptors from the scoring scale and the anchor responses to confirm the rationale behind each response's score. Starting on the sixth day that a rater was working on a task, DRC used validity responses to continue monitoring rater performance. DRC seeded the validity responses into the operational scoring so that the raters did not know which responses were operational and which were validity responses. Reports generated daily compared the scores that each rater assigned to the "true" score for each validity response. When a rater was working on a task, DRC seeded the validity responses in random order into the rater's queue for scoring. Given enough time, every rater working on a task would score every validity response for that task, but the order in which the raters would see the validity responses would differ.

Handling Unusual Responses

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

- **Scoring questions.** If a rater had questions about the application of the scoring guidelines to a response (e.g., if they were uncertain as to the proper score that they should assign), the rater forwarded the response to their team leader for assistance. The team leader then reviewed the response with the rater and assigned the proper score. If the rater needed further clarifications, the team leader worked with the rater to review scoring guidelines.

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

Remote Scoring Procedures due to the COVID-19 Pandemic

Prior to 2020, DRC's handscoring centers managed all WIDA handscoring. In 2020, due to the COVID-19 pandemic, DRC shifted from site-based handscoring to remote handscoring to continue meeting all the handscoring deadlines. All WIDA handscoring continued to be remote

in 2021. DRC designed the remote scoring to very closely emulate the work carried out in the physical scoring locations. The platform, content, and expectations for quality remained the same. Using a variety of modes of interactive technology (i.e., web screen sharing, webcast, video chat, and chat), DRC conducted rater training and discussions live (virtually). DRC equipped scoring leaders with a variety of tools to ensure that every rater was successful in understanding and applying scoring criteria to student responses.

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

- **System tools—scoring, training, chat.** ScoreBoard is DRC's secure, web-based scoring application that is designed to be used in a distributed environment. The platform is used within DRC's scoring centers and in remote locations (e.g., in a rater's home). Integrated training resources provide the capability to securely maintain digital training materials within the scoring platform itself.
- DRC conducted live, interactive rater training using the Moodle Learning Management System, which mirrored aspects of the scoring room and provided a versatile platform for training. It also served as a place to share files of important documents, including daily scoring statistics and platform user guides. Through embedded communication tools, scoring directors, assistant scoring directors, and team leaders facilitated group and one-on-one training sessions and discussions using audio and video.
- To facilitate instant communication between supervisors and raters, DRC utilized a chat tool called Zulip in conjunction with ScoreBoard and Moodle. Zulip provided a tool for raters to directly ask supervisors questions about responses and allowed supervisors to direct individuals or groups of raters to join Moodle training rooms for important discussions and retraining.
- **Security.** Security is essential to the handscoring process. When users logged into ScoreBoard, they were required to read and accept the security policy before they were allowed to access the project. DRC also required raters to read and sign nondisclosure agreements. During training and large-group discussions, trainers continuously emphasized what security means, the importance of maintaining security, and how all

staff accomplish this. In the remote environment, DRC could give these security reminders daily. DRC requires raters working remotely to work in a private environment away from other people (including family members). Raters working in ScoreBoard were not allowed to print from their computers in order to protect the security of the student responses, test questions, and training materials. Restrictions built into ScoreBoard defined the hours during the day that raters were able to log into the system, ensuring that raters were only scoring responses while supervisors were in place to monitor handscoring and answer any questions.

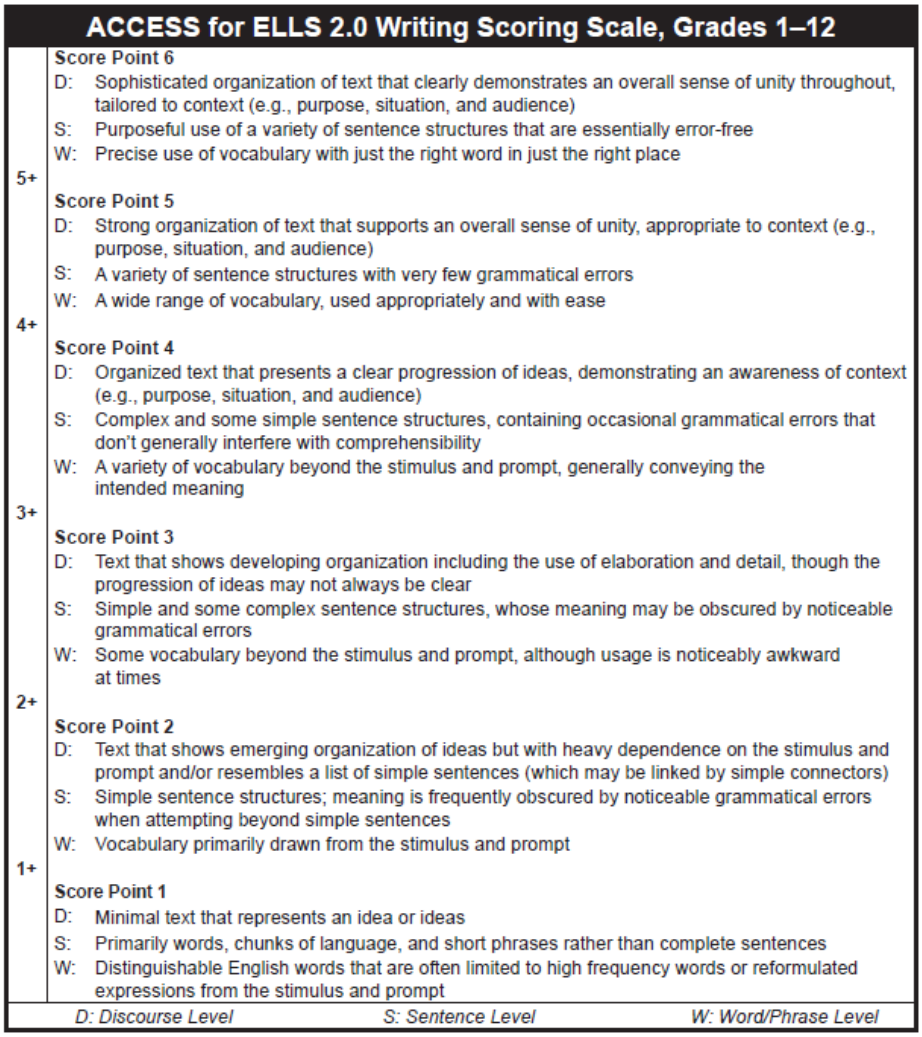
- **Rater training with Moodle.** DRC conducted rater training remotely as an interactive, comprehensive, hands-on experience. For Writing training, scoring directors trained groups of raters by screensharing PDFs of training materials. Raters individually viewed each training example, with supervisors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

4.3 Writing Scoring Scale

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

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

Figure 20 shows the Writing Scoring Scale.



ACCESS for ELLS 2.0 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*

Figure 20. Writing Scoring Scale

When assigning a score, a rater makes an initial judgment about which whole score point (1–6) best describes a response and then determines whether the three descriptors for that whole score point suit that response. If all three descriptors suit the response, the rater assigns the score associated with that score point (e.g., if all three descriptors for score point 3 are appropriate, the rater would assign a score of 3). However, if there is clear evidence that one or two descriptors from an adjacent score point are a better fit, the rater would assign a plus score between the two applicable whole score points (e.g., if two descriptors for score point 3 seem to fit, but one descriptor for score point 4 is a better fit than the associated descriptor for score point 3, the rater would assign a score of 3+).

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

Nonscorable: The response is blank; consists only of verbatim copied text; consists only of text that is completely off task; or is entirely in a language other than English; or appears to have been plagiarized from an outside source during testing. More information on how plagiarized responses are handled by DRC is provided in Section Scoring Performance-Based Tasks: Writing and Speaking, *Handling Unusual Responses*, *Alerts* above.

Completely off-task response: The entire response shows no understanding of or interaction with the prompt. It may be a memorized, previously practiced response or appear to answer another, unrelated prompt. A response that is entirely off task is nonscorable.

Completely off-topic response: The entire response shows a misinterpretation or misunderstanding of the prompt. An off-topic response is related to the prompt but does not seem to address it as intended. However, the response is clearly not a memorized, previously practiced response. Raters score these responses in their entirety using the scoring scale; however, the maximum score for a completely off-topic response is 2+.

Partially off-task response: The response contains both off-task and on-task writing. Raters score these responses by ignoring the off-task portion (which may be memorized and previously practiced) and scoring only the on-task portion using the scoring scale.

Partially off-topic response: The response contains both off-topic and on-topic writing (i.e., a portion of the response shows a misinterpretation or misunderstanding of the prompt). Raters score these responses in their entirety using the scoring scale.

Each student responds to two Writing tasks. One rater assigns a score to each student's response for each task. To calculate a student's total raw score by task, the scores that the raters assigned are converted to whole numbers ranging from 0 to 9, as shown in Table 14. The Writing scoring scale was designed to go up to score point 6. However, we did not have enough responses to estimate the rating scale parameters at each of the 5, 5+, and 6 score points in the empirical data. Therefore, these score points were collapsed into one category for psychometric purposes. Students' scores are then added across tasks, resulting in a total raw score that ranges from 0 to 18.

Table 14. Rating to Raw Score Conversion (Writing)

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

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

4.4 Speaking Scoring Scale

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

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

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

Figure 21 shows the Speaking Scoring Scale.

ACCESS for ELLs 2.0 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)

Figure 21. Speaking Scoring Scale

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

Each student responds to three (or six) Speaking tasks, depending upon how the test engine routes the student. A single rater assigns a score to each of those responses, as shown in Table 15. To calculate a total raw score, the scores are then summed, based on the following guidelines:

- For tasks targeting language elicitation at PL 1, there are only three possible score points: *No Response*, *Attempted*, and *Adequate and Above*. This is the case because appropriate responses to PL 1 tasks are single words and short chunks of language, so it is not possible to reliably distinguish between *Adequate*, *Strong*, and *Exemplary* performances.
- For tasks targeting language elicitation at PL 3 and PL 5, each task can be scored on the entire breadth of the scale.
- Each student routed to Tier Pre-A responds to three PL 1 Speaking tasks. Thus, for students in this tier, the total raw score can range from 0 to 6.
- Students routed to Tier A respond to six Speaking tasks, three at PL 1 and three at PL 3. For students in this tier, the total raw score can range from 0 to 18.
- When scoring students' responses to Speaking tasks included in Tier B/C, six points are added to the total raw score, representing a score of *Adequate and Above* for three tasks targeting language at PL 1. Though a Tier B/C student would not be administered any tasks targeting the PL 1 level, it is assumed that a student who had been routed to Tier B/C would easily achieve a score of *Adequate and Above* on these tasks. Thus, for a student routed to Tier B/C, the total raw score can range from 6 to 30.

Table 15. Score to Raw Score Conversion (Speaking)

Score	Raw score
No Response (R, F, I, or K)*	0
Attempted	1
Adequate/Adequate and Above	2
Strong	3
Exemplary	4

*R = Dead air or white noise; F = Foreign language response; I = Nonscorable utterance; K = suspected plagiarism.

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

5. Summary of Score Reports

5.1 Individual Student Report

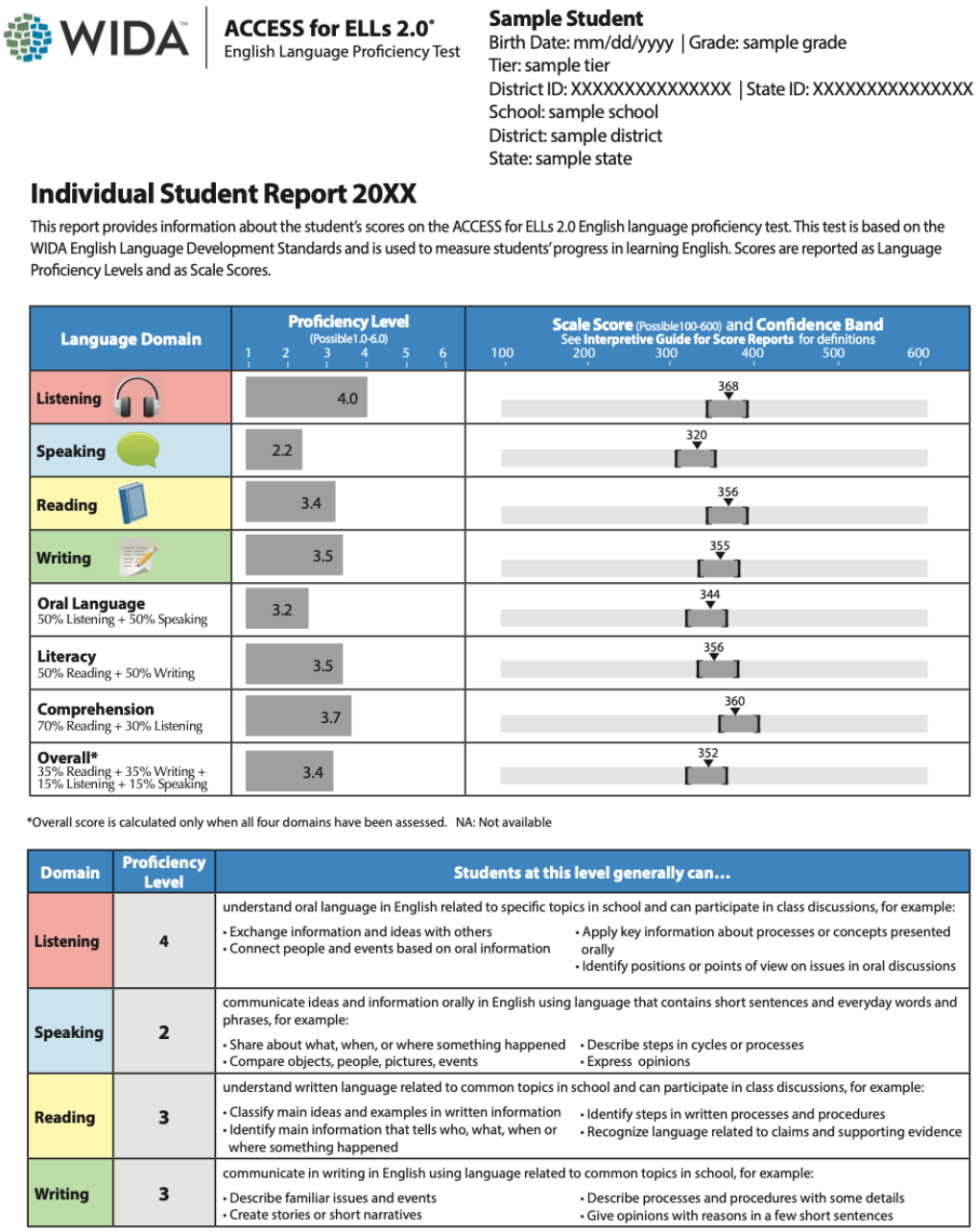
Score reports (district, school, and student level reports) are made available in WIDA Assessment Management System (AMS) as soon as they are available for each state, and WIDA ships printed reports to school districts and schools at the same time or shortly thereafter. Score reports are available for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

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

WIDA offers several online resources to help communicate test score information to educators, families, and students. (See ACCESS for ELLs Score and Reports https://wida.wisc.edu/assess/access/scores-reports; Family Engagement https://wida.wisc.edu/teach/learners/engagement). WIDA also provides a post-testing Q & A webinar about score interpretation (https://portal.wida.us/webinar/detail/702b69ef-0265-eb11-a2dd-0050568beee8).

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

The Individual Student Report (Figure 22) contains detailed information about the performance of a single student in Grades K–12. Its primary users are students, parents/guardians, teachers, and school teams. It provides information about one indicator of a student's English language proficiency: the language needed to access content and succeed in school.



WIDA | ACCESS for ELLs 2.0® English Language Proficiency Test

Sample Student
Birth Date: mm/dd/yyyy | Grade: sample grade
Tier: sample tier
District ID: XXXXXXXXXXXXXX | State ID: XXXXXXXXXXXXXX
School: sample school
District: sample district
State: sample state

Individual Student Report 20XX

This report provides information about the student's scores on the ACCESS for ELLs 2.0 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 (Possible1.0-6.0) 1 2 3 4 5 6	Scale Score (Possible100-600) and Confidence Band See Interpretive Guide for Score Reports for definitions 100 200 300 400 500 600
Listening	4.0	368
Speaking	2.2	320
Reading	3.4	356
Writing	3.5	355
Oral Language 50% Listening + 50% Speaking	3.2	344
Literacy 50% Reading + 50% Writing	3.5	356
Comprehension 70% Reading + 30% Listening	3.7	360
Overall* 35% Reading + 35% Writing + 15% Listening + 15% Speaking	3.4	352

*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	4	understand oral language in English related to specific topics in school and can participate in class discussions, for example: • Exchange information and ideas with others • Connect people and events based on oral information • Apply key information about processes or concepts presented orally • Identify positions or points of view on issues in oral discussions
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, events • Describe steps in cycles or processes • Express opinions
Reading	3	understand written language related to common topics in school and can participate in class discussions, for example: • Classify main ideas and examples in written information • Identify main information that tells who, what, when or where something happened • Identify steps in written processes and procedures • Recognize language related to claims and supporting evidence
Writing	3	communicate in writing in English using language related to common topics in school, for example: • Describe familiar issues and events • Create stories or short narratives • Describe processes and procedures with some details • Give opinions with reasons in a few short sentences

Figure 22. Individual Student Report

The Individual Student Report includes four language domain scores (Listening, Speaking, Reading, and Writing) and four language domain composite scores (Oral Language, Literacy, Comprehension, and Overall), as shown in the first table of the score report. In the first column of the last four rows of that table, test users can see how WIDA uses a student's domain scores to calculate each composite score (e.g., for Oral Language, WIDA calculates the composite score based on a student's performance on the Listening and Speaking tests, with scores on each of

those tests contributing equally to the composite score). For students who are unable to complete all four domains due to their disabilities, WIDA provides states methods to compute alternative composite scores based on their available domain scores upon request (Sahakyan, 2020).

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

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

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

The second table in the Individual Student Report provides information about the student's proficiency levels expressed as whole numbers. The third column of the table describes what that student should generally be able to do in each of the four language domains, given his or her level of proficiency. For example, as shown in Figure 22, 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 native language will most likely acquire literacy in English at a quicker pace than students who do not.
- The Overall score is helpful as a summary of other scores and is used because a single number may be needed for reference. However, it is important to remember that it is compensatory, averaged using weights; a particularly high score in one domain may effectively offset a low score in another domain and vice versa. Similar Overall scores can mask very different performances on individual tests.
- No single scale score or language proficiency level, including the Overall score (composite), should be used as the sole determiner for making decisions regarding a student's English language proficiency. School work and local assessment throughout the school year also provide evidence of a student's English language development.
- Scale scores can be used to make comparisons across grade levels, but not across domains. Each domain has its own score scale, so scale scores should not be used for comparing performance across domains. For example, a scale score of 350 in Listening at grade 3 is not equivalent to a scale score of 350 in Speaking at grade 3. For performance comparisons across domains, proficiency levels should be used.

- Either scale scores or proficiency levels can be used to compare test performance from different years, although it is easier to see changes when examining scale scores.

For detailed information about score reports, please refer to the Interpretive Guide.

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.

Part 2:

Technical Results

Contents

1. Annual Test Results

This section of the report provides an overview of students' participation, the distribution of students' scale scores, and the distribution of students' proficiency levels to see student performance of the ACCESS 601 administration. Results are presented, where appropriate, by grade-level cluster, grade, and tier (for Writing and Speaking), and also by state, by gender, and by race and ethnicity.

Following the approach of the U.S. Census Bureau (https://www.census.gov/topics/population/race/about.html), ethnicity is a binary category (Hispanic or non- Hispanic), with five categories for race (American Indian/Alaskan Native, Asian, Black/African American, Pacific Islander/Hawaiian, and White) that are not mutually exclusive. Thus, for example, Student A may be labeled as Hispanic for ethnicity and Asian for race, while Student B may be labeled as non-Hispanic for ethnicity and both American Indian/Alaskan Native and Black/African American for race. Students who are labeled Hispanic are included in the Hispanic (of any race) category, regardless of how many racial categories they are included in. Students who are identified in one racial category (e.g., Asian) who have not been identified as Hispanic are identified in only one racial category; if they are identified in more than one racial category and have not been identified as Hispanic, they are labeled non-Hispanic multiracial.

A subset of students was included in the descriptions of student participation and performance but were excluded from subsequent analyses, namely those students who were flagged as potentially having experienced test interruptions. Using telemetry data, WIDA selected three variables that might potentially indicate interruption (that is, testing experiences that are outside of regular testing experiences). The interruption indicators WIDA used are (1) longer than expected testing time, (2) number of appearances (e.g., more than one) of test items, and (3) number of log-ins. Records were flagged if they fell outside of established criteria for any of these three indicators. WIDA included students whose records were flagged as interrupted in the

tables that describe participation in the assessment but excluded them from all subsequent analyses. Table 1.1 summarizes the numbers of students excluded from these analyses. On average, 3% to 9% of students were excluded in each cluster and domain.

Table 1.1

Students Excluded from Analysis Due to Test Interruptions by Domain and Cluster

Domain	Cluster	No. of Excluded Students	Total Students	Percent
Listening	1	14,639	227,996	6.42%
	2-3	30,973	446,174	6.94%
	4-5	26,173	391,153	6.69%
	6-8	40,142	454,273	8.84%
	9-12	32,218	478,541	6.73%
	Total	144,145	1,998,137	7.21%
Reading	1	8,955	227,996	3.93%
	2-3	22,652	446,174	5.08%
	4-5	27,789	391,153	7.10%
	6-8	36,948	454,273	8.13%
	9-12	37,846	478,541	7.91%
	Total	134,190	1,998,137	6.72%
Writing	1	n/a	227,996	n/a
	2-3	n/a	446,174	n/a
	4-5	18,256	391,153	4.67%
	6-8	25,688	454,273	5.65%
	9-12	21,684	478,541	4.53%
	Total	65,628	1,323,967	4.96%

Domain	Cluster	No. of Excluded Students	Total Students	Percent
Speaking	1	17,356	227,996	7.61%
	2-3	33,074	446,174	7.41%
	4-5	29,243	391,153	7.48%
	6-8	43,363	454,273	9.55%
	9-12	36,578	478,541	7.64%
	Total	159,614	1,998,137	7.99%

1.1 Participation

Participation in ACCESS Online is shown in three ways: by grade-level cluster, by grade, and, for Writing and Speaking only, by tier.

1.1.1 Grade-Level Cluster

Table 1.1.1.1 shows participation across the 40 WIDA states and U.S. territories that participated in the ACCESS Online operational testing program in 2022–2023 by grade-level cluster. The 40 rows show the number of students in that grade-level cluster who took the test by state, and the final row shows the total number of participants across all 40 states and U.S. territories. The state with the largest number of students was Illinois. The state/territory with the smallest number of participants was VI. The biggest cluster was Grade 9-12. The abbreviations are as follows: DC, District of Columbia; DD, Department of Defense Education Activity; MP, Northern Mariana Islands; BI, Bureau of Indian Education, and VI, Virgin Islands.

Table 1.1.1.1

Participation by Cluster by State, S601 Online

State	Cluster					Total
	1	2–3	4–5	6–8	9–12	
AK	849	1936	1932	2812	2881	10410
AL	4250	7710	7351	9201	8038	36550
BI	274	605	713	705	422	2719
CO	9916	18734	15055	17099	18559	79363
DC	1108	2110	1663	1965	1788	8634
DD	785	1534	1223	1137	736	5415
DE	1604	3142	2889	3288	3418	14341
GA	15958	30457	26158	29048	25636	127257
HI	1549	3332	3358	4001	3617	15857
ID	1734	3960	3348	3628	4007	16677
IL	24326	49606	44575	54946	50561	224014
IN	8539	17223	15798	18794	18051	78405
KY	4667	8725	6812	6965	8395	35564
MA	12581	22934	16813	17906	23516	93750
MD	11610	21939	18700	19474	21956	93679
ME	521	1143	950	1130	1428	5172
MI	8604	17503	15888	18745	24561	85301
MN	8043	16483	13763	13821	14975	67085
MO	4288	7891	6771	6770	6893	32613
MP	108	226	282	410	395	1421
MT	250	675	681	772	601	2979

State	Cluster					Total
	1	2–3	4–5	6–8	9–12	
NC	12999	25525	25680	32249	29977	126430
ND	412	842	716	775	808	3553
NH	588	1009	868	948	1105	4518
NJ	13762	24918	19930	22241	25926	106777
NM	4543	8641	8950	13472	14579	50185
NV	6081	12411	10843	12932	15375	57642
OK	6449	13155	11884	14927	13758	60173
PA	8429	15682	14247	18105	21612	78075
RI	1567	3119	2719	3721	4725	15851
SC	4561	8953	8219	10886	13302	45921
SD	786	1447	1080	1257	1317	5887
TN	8017	13718	11124	11789	13603	58251
UT	4618	10271	11047	15120	13298	54354
VA	13457	27655	22965	22445	26408	112930
VI	62	187	210	312	271	1042
VT	162	326	281	331	351	1451
WA	14799	29778	24950	27384	28526	125437
WI	4890	10141	10261	12292	12529	50113
WY	250	528	456	470	637	2341
Total	227996	446174	391153	454273	478541	1998137

Table 1.1.1.2 shows participation by grade-level cluster by gender across all 40 states and U.S. territories combined, while Table 1.1.1.3 shows participation by grade-level cluster by ethnicity across all 40 states and U.S. territories. The gender ratio was generally 40% female, 45% male and 15% missing gender information in Clusters. About 64-67% of participants were Hispanic across all clusters.

Table 1.1.1.2

Participation by Cluster by Gender, S601 Online

Cluster		Gender			Total
		F	M	Missing	
1	Count	93564	102749	31683	227996
	% within Cluster	41.03%	45.06%	13.89%	100%
2–3	Count	181754	201362	63058	446174
	% within Cluster	40.73%	45.13%	14.13%	100%
4–5	Count	153363	180047	57743	391153
	% within Cluster	39.20%	46.02%	14.76%	100%
6–8	Count	172807	213561	67905	454273
	% within Cluster	38.04%	47.01%	14.94%	100%
9–12	Count	180143	228362	70036	478541
	% within Cluster	37.64%	47.72%	14.63%	100%
Total	Count	781631	926081	290425	1998137
	% within Cluster	39.12%	46.35%	14.53%	100%

Table 1.1.1.3

Participation by Cluster by Ethnicity, S601 Online

Cluster		Ethnicity			Total
		Hispanic	Non-Hispanic	Unknown	
1	Count	146560	66535	14901	227996
	% within Cluster	64.28%	29.18%	6.53%	100%
2–3	Count	286797	130658	28719	446174
	% within Cluster	64.27%	29.28%	6.43%	100%
4–5	Count	255786	103761	31606	391153
	% within Cluster	65.39%	26.52%	8.08%	100%
6–8	Count	307485	104418	42370	454273
	% within Cluster	67.68%	22.98%	9.32%	100%
9–12	Count	320521	110050	47970	478541
	% within Cluster	66.97%	22.99%	10.02%	100%
Total	Count	1317149	515422	165566	1998137
	% within Cluster	65.92%	25.80%	8.29%	100%

Table 1.1.1.4 shows participation by grade-level cluster and tier for all Writing and Speaking forms. In the Writing domain, Cluster 1 had a higher percentage of Tier A than Tier B/C, while in other Clusters, percentages of Tier A became smaller. In the Speaking domain, percentages of Tier A remained smaller than Tier B/C for all clusters. Pre-A counts in Speaking were relatively small.

Table 1.1.1.4

Participation by Cluster by Tier by Domain, S601 Online

Cluster			Domain	
			Writing	Speaking
1	Tier	Pre-A	-	14198
		A	205745	109422
		BC	22218	104373
	Total		227963	227993
2–3	Tier	Pre-A	-	21285
		A	137309	130017
		BC	308803	294868
	Total		446112	446170
4–5	Tier	Pre-A	-	8714
		A	89609	61481
		BC	301544	320958
	Total		391153	391153
6–8	Tier	Pre-A	-	16901
		A	181387	104181
		BC	272877	333184
	Total		454264	454266
9–12	Tier	Pre-A	-	34677
		A	179699	202520
		BC	298784	241304

Cluster		Domain	
		Writing	Speaking
	Total	478483	478501

1.1.2 Grade

This section provides tables parallel to those in the previous section but broken out by grade rather than by grade-level cluster. Table 1.1.2.1 shows student counts by grade and state. The largest grade was 2nd grade, and the smallest was 12th grade. Table 1.1.2.4 presents the percentages between Tier A and B/C and indicates that most grades showed higher counts in tier B/C forms except in Writing grade 1.

Table 1.1.2.1

Participation by Grade by State, S601 Online

State	Grade												
	1	2	3	4	5	6	7	8	9	10	11	12	Total
AK	849	891	1045	1044	888	906	934	972	901	822	617	541	**10410**
AL	4250	3922	3788	3844	3507	3017	3126	3058	3249	2353	1350	1086	**36550**
BI	274	288	317	331	382	307	211	187	155	112	96	59	**2719**
CO	9916	9669	9065	8223	6832	5645	5866	5588	5991	5273	3952	3343	**79363**
DC	1108	1052	1058	958	705	619	709	637	854	498	230	206	**8634**
DD	785	752	782	684	539	425	369	343	251	221	162	102	**5415**
DE	1604	1550	1592	1558	1331	1078	1160	1050	1441	986	555	436	**14341**
GA	15958	15185	15272	14865	11293	9301	10188	9559	10340	7245	4588	3463	**127257**
HI	1549	1576	1756	1828	1530	1398	1483	1120	1438	927	620	632	**15857**
ID	1734	1999	1961	1941	1407	1289	1220	1119	1360	1103	855	689	**16677**
IL	24326	25478	24128	23852	20723	18810	18387	17749	17839	14631	10199	7892	**224014**

State	Grade												
	1	2	3	4	5	6	7	8	9	10	11	12	Total
IN	8539	8406	8817	8478	7320	6252	6309	6233	6305	5031	3671	3044	**78405**
KY	4667	4342	4383	4019	2793	2446	2306	2213	3050	2347	1680	1318	**35564**
MA	12581	12008	10926	9827	6986	5945	6177	5784	7746	6719	4943	4108	**93750**
MD	11610	11457	10482	10427	8273	6922	6731	5821	9770	6347	2976	2863	**93679**
ME	521	594	549	542	408	388	386	356	369	405	361	293	**5172**
MI	8604	8756	8747	8687	7201	6353	6391	6001	7190	6673	5386	5312	**85301**
MN	8043	8440	8043	7713	6050	4886	4680	4255	4773	4184	3376	2642	**67085**
MO	4288	3908	3983	3801	2970	2380	2285	2105	2197	1994	1489	1213	**32613**
MP	108	109	117	126	156	154	136	120	114	138	106	37	**1421**
MT	250	320	355	373	308	265	236	271	218	185	108	90	**2979**
NC	12999	12592	12933	13565	12115	10617	10903	10729	12801	8338	5148	3690	**126430**
ND	412	428	414	352	364	227	282	266	281	215	153	159	**3553**
NH	588	504	505	505	363	326	336	286	356	327	213	209	**4518**
NJ	13762	12909	12009	11108	8822	7663	7603	6975	7992	7318	5716	4900	**106777**
NM	4543	4694	3947	4411	4539	4224	4664	4584	5325	4189	2832	2233	**50185**
NV	6081	5930	6481	5899	4944	4508	4300	4124	4247	4151	3767	3210	**57642**
OK	6449	6535	6620	6462	5422	4979	5074	4874	5013	4057	2740	1948	**60173**
PA	8429	8037	7645	7636	6611	6047	6168	5890	6826	5803	4764	4219	**78075**
RI	1567	1493	1626	1449	1270	1199	1288	1234	1519	1347	972	887	**15851**
SC	4561	4463	4490	4399	3820	3636	3858	3392	4851	3602	2794	2055	**45921**
SD	786	686	761	590	490	415	438	404	503	397	241	176	**5887**

State	Grade												
	1	2	3	4	5	6	7	8	9	10	11	12	Total
TN	8017	6991	6727	6268	4856	4164	3949	3676	4838	3965	2733	2067	**58251**
UT	4618	4993	5278	5747	5300	5116	4987	5017	4657	3883	2878	1880	**54354**
VA	13457	13657	13998	12883	10082	8091	7470	6884	9596	7676	5216	3920	**112930**
VI	62	85	102	106	104	80	115	117	106	82	48	35	**1042**
VT	162	157	169	147	134	108	111	112	116	91	81	63	**1451**
WA	14799	14634	15144	13674	11276	9867	9006	8511	8271	7689	6622	5944	**125437**
WI	4890	4906	5235	5317	4944	4152	4213	3927	3932	3536	2812	2249	**50113**
WY	250	276	252	245	211	147	159	164	153	181	136	167	**2341**
Total	**227996**	**224672**	**221502**	**213884**	**177269**	**154352**	**154214**	**145707**	**166934**	**135041**	**97186**	**79380**	**1998137**

Table 1.1.2.2

Participation by Grade by Gender, S601 Online

Grade		Gender			Total
		F	M	Missing	
1	Count	93564	102749	31683	227996
	% within Grade	41.04%	45.07%	13.90%	100.00%
2	Count	92912	100914	30846	224672
	% within Grade	41.35%	44.92%	13.73%	100.00%
3	Count	88842	100448	32212	221502
	% within Grade	40.11%	45.35%	14.54%	100.00%
4	Count	84854	97928	31102	213884
	% within Grade	39.67%	45.79%	14.54%	100.00%
5	Count	68509	82119	26641	177269
	% within Grade	38.65%	46.32%	15.03%	100.00%
6	Count	58830	72156	23366	154352
	% within Grade	38.11%	46.75%	15.14%	100.00%
7	Count	58779	72790	22645	154214
	% within Grade	38.12%	47.20%	14.68%	100.00%

8	Count	55198	68615	21894	145707
	% within Grade	37.88%	47.09%	15.03%	100.00%
9	Count	61951	79894	25089	166934
	% within Grade	37.11%	47.86%	15.03%	100.00%
10	Count	50060	65596	19385	135041
	% within Grade	37.07%	48.57%	14.35%	100.00%
11	Count	37103	45777	14306	97186
	% within Grade	38.18%	47.10%	14.72%	100.00%
12	Count	31029	37095	11256	79380
	% within Grade	39.09%	46.73%	14.18%	100.00%
Total	Count	781631	926081	290425	1998137
	% within Grade	39.11%	46.34%	14.53%	100.00%

Table 1.1.2.3

Participation by Grade by Ethnicity, S601 Online

Grade		Ethnicity			Total
		Hispanic	Non-Hispanic	Unknown	
1	Count	146560	66535	14901	227996
	% within Grade	64.28%	29.18%	6.54%	100.00%
2	Count	144509	65598	14565	224672
	% within Grade	64.32%	29.20%	6.48%	100.00%
3	Count	142288	65060	14154	221502
	% within Grade	64.24%	29.37%	6.39%	100.00%
4	Count	138101	59043	16740	213884
	% within Grade	64.57%	27.61%	7.83%	100.00%
5	Count	117685	44718	14866	177269
	% within Grade	66.39%	25.23%	8.39%	100.00%
6	Count	104298	35834	14220	154352
	% within Grade	67.57%	23.22%	9.21%	100.00%
7	Count	103987	35773	14454	154214
	% within Grade	67.43%	23.20%	9.37%	100.00%
	Count	99200	32811	13696	145707

Grade		Ethnicity			Total
		Hispanic	Non-Hispanic	Unknown	
	% within Grade	68.08%	22.52%	9.40%	100.00%
9	Count	114363	35413	17158	166934
	% within Grade	68.51%	21.21%	10.28%	100.00%
10	Count	92071	29990	12980	135041
	% within Grade	68.18%	22.21%	9.61%	100.00%
11	Count	63513	23664	10009	97186
	% within Grade	65.35%	24.35%	10.30%	100.00%
12	Count	50574	20983	7823	79380
	% within Grade	63.71%	26.43%	9.86%	100.00%
Total	Count	1317149	515422	165566	1998137
	% within Grade	65.91%	25.79%	8.26%	100.00%

Table 1.1.2.4

Participation by Grade by Tier by Domain, S601 Online

Grade			Domain	
			Writing	Speaking
1	Tier	Pre-A	-	14198
		A	205745	109422
		BC	22218	104373
	Total		227963	227993
2	Tier	Pre-A	-	6357
		A	75034	65517
		BC	149601	152795

Grade			Domain	
			Writing	Speaking
	Total		224635	224669
3	Tier	Pre-A	-	14928
		A	62275	64500
		BC	159202	142073
	Total		221477	221501
4	Tier	Pre-A	-	2460
		A	44841	34271
		BC	169043	177153
	Total		213884	213884
5	Tier	Pre-A	-	6254
		A	44768	27210
		BC	132501	143805
	Total		177269	177269
6	Tier	Pre-A	-	2987
		A	51372	31440
		BC	102978	119924
	Total		154350	154351
7	Tier	Pre-A	-	5401
		A	63278	25434
		BC	90933	123378

Grade			Domain	
			Writing	Speaking
	Total		154211	154213
8	Tier	Pre-A	-	8513
		A	66737	47307
		BC	78966	89882
	Total		145703	145702
9	Tier	Pre-A	-	9089
		A	69526	91436
		BC	97387	66394
	Total		166913	166919
10	Tier	Pre-A	-	9431
		A	51619	55467
		BC	83409	70133
	Total		135028	135031
11	Tier	Pre-A	-	8394
		A	33406	20205
		BC	63769	68581
	Total		97175	97180
12	Tier	Pre-A	-	7763
		A	25148	35412
		BC	54219	36196

Grade		Domain	
		Writing	Speaking
	Total	79367	79371

1.2 Scale Score Results

This section provides information on students' scale score results.

1.2.1 Mean Scale Score Across Domain and Composite Score by Cluster

This section shows mean (average) scale scores by grade-level cluster across the eight scores awarded, first for the four domains (Listening, Reading, Writing, and Speaking) and then for the four composites (Oral Language, Literacy, Comprehension, and Overall Composite). The mean scale scores are expected to increase as grade increases, as ACCESS is vertically scaled, but there is also an intersection between this principle and the population of test-takers.

In this section, under each average, the number of students in each group is also given. In Table 1.2.1.1, the order of average scale scores among single domains in descending order were Listening, Reading, Writing, and then Speaking except clusters 1 and 2-3. Cluster 4–5 showed the highest average scale score in Listening domain across all clusters, and scores dropped in Cluster 6–8.

Table 1.2.1.2 demonstrates that groups made up of female students performed better than groups of male students in general. Table 1.2.1.3 presents scale score performance by ethnic groups. The top three performing ethnic groups were Asian students, White students, and multiracial students in most domains and clusters. Additional tables show this information by gender, and by race and ethnicity.

Table 1.2.1.1

Mean Scale Scores by Cluster, S601 Online

Cluster		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Mean	303.26	283.68	237.99	240.09	271.81	260.9	289.56	263.97
	N	213228	218896	227850	210557	198705	218829	206299	192851
2–3	Mean	321.01	322.89	286.5	271.28	296.3	304.7	322.35	301.97
	N	415059	423327	445905	412991	388000	423167	397807	373332
4–5	Mean	403.01	346.97	325.66	316.14	359.87	336.26	363.9	343.24
	N	364830	363103	372713	361784	340995	349848	343831	313679
6–8	Mean	390.98	348.79	315.94	310.16	350.77	332.36	361.66	337.68
	N	413790	416954	428278	410683	380893	400485	389590	350963
9–12	Mean	387.95	377.76	341.66	303.88	346.02	359.76	380.95	355.26
	N	445815	440071	456246	441386	415651	424984	416896	381032

Table 1.2.1.2

Mean Scale Scores by Gender, S601 Online

Cluster	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	F	Mean	306.1	283.68	241.93	245.61	275.93	262.85	290.39	266.47
		N	88270	89959	93501	87395	83081	89935	85487	80693
	M	Mean	300.09	283.94	234.85	235.33	267.9	259.49	288.82	261.86
		N	95972	99000	102669	94755	89264	98960	93128	86845
	Missing	Mean	305.12	282.81	236.5	239.02	272.06	259.7	289.5	263.25
		N	28986	29937	31680	28407	26360	29934	27684	25313
2–3	F	Mean	322.01	323.76	292	276.88	299.59	307.87	323.24	305.09
		N	170376	172454	181659	169893	160508	172410	163174	154322
	M	Mean	319.62	322.24	282.14	266.9	293.5	302.22	321.52	299.47
		N	187537	192039	201220	186570	175401	191948	180609	169457
	Missing	Mean	322.57	322.5	284.57	268.89	295.56	303.48	322.45	300.8
		N	57146	58834	63026	56528	52091	58809	54024	49553
4–5	F	Mean	402.29	347.58	332.01	319.2	361	339.73	364.06	345.9
		N	143943	142330	146146	142586	135036	137186	135498	124068
	M	Mean	402.98	346.32	321.52	314.09	358.85	333.87	363.44	341.28
		N	168153	168283	171890	166947	157452	162284	159419	145622
		Mean	405.04	347.43	321.69	314.33	360.03	334.5	364.97	342.26

Cluster	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
	Missing	N	52734	52490	54677	52251	48507	50378	48914	43989
6–8	F	Mean	389.27	350.17	319.8	311.5	350.48	334.99	362.04	339.26
		N	159018	158872	162975	156782	146610	152673	149728	135122
	M	Mean	392.15	347.87	313.39	309.9	351.3	330.61	361.38	336.71
		N	194398	196835	201769	194000	179599	189145	183577	165750
	Missing	Mean	391.7	348.17	314.1	307.46	349.76	331.13	361.51	336.6
		N	60374	61247	63534	59901	54684	58667	56285	50091
9–12	F	Mean	387.45	380.01	345.14	307.05	347.31	362.66	382.37	357.61
		N	168874	165780	171652	166676	157833	160142	157970	144737
	M	Mean	387.93	375.85	339.38	301.67	344.95	357.67	379.61	353.55
		N	212351	210611	217951	211101	198379	203473	199136	182243
	Missing	Mean	389.33	378.23	340.15	302.92	346.16	359.17	381.62	354.74
		N	64590	63680	66643	63609	59439	61369	59790	54052

Table 1.2.1.3

Mean Scale Scores by Ethnicity, S601 Online

Cluster	**Ethnicity**		**Listening**	**Reading**	**Writing**	**Speaking**	**Oral**	**Literacy**	**Compre-hension**	**Overall**
1	Hispanic (of any Race)	Mean	299.31	279.91	231.54	235.77	267.68	255.8	285.72	259.17
		N	137312	141152	146488	135707	128221	141101	133176	124716
	Non-Hispanic American Indian	Mean	307.6	282.27	232.09	241.77	274.82	257.52	290.25	263.11
		N	1355	1400	1464	1337	1253	1399	1298	1201
	Non-Hispanic Asian	Mean	318.7	299.19	263.12	254.17	286.72	281.36	305.13	282.95
		N	26672	27231	28406	26300	24868	27228	25776	24113
	Non-Hispanic Black	Mean	307.61	288.16	245.33	256.1	281.9	266.84	294.1	271.05
		N	10534	10822	11363	10350	9682	10815	10095	9305
	Non-Hispanic Multiracial	Mean	315.03	290.03	249.26	250.36	282.73	269.65	297.61	273.31
		N	1038	1071	1103	1028	974	1071	1015	957
	Non-Hispanic Pacific Islander	Mean	291.11	279.7	236.11	233.28	262.36	258	283.21	259.55
		N	1662	1729	1814	1681	1557	1728	1599	1511
	Non-Hispanic White	Mean	313.95	288.66	250.2	248.6	281.51	269.48	296.24	272.93
		N	20857	21254	22321	20492	19314	21250	20016	18611
	Unknown	Mean	293.36	280.76	229.53	230.97	262.02	254.99	284.45	256.5
		N	13798	14237	14891	13662	12836	14237	13324	12437
2–3	Hispanic (of any Race)	Mean	316.34	319.57	281.95	268.21	292.42	300.77	318.62	298.06
		N	267421	272849	286647	266314	250534	272740	256762	241404
	Non-Hispanic American Indian	Mean	323.61	319.11	282.71	268.75	296.18	301.12	320.59	299.53
		N	2886	2982	3119	2895	2700	2978	2774	2597
	Non-Hispanic Asian	Mean	339.8	336.86	305.55	283.21	311.78	321.36	337.84	318.4
		N	50732	51606	54240	50391	47556	51595	48746	45891

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
	Non-Hispanic Black	Mean	327.61	326.04	290.95	284.67	306.44	308.55	326.55	307.8
		N	21519	22028	23340	21374	19916	22018	20532	19105
	Non-Hispanic Multiracial	Mean	336.19	331.2	296.08	282.72	309.8	313.67	332.81	312.4
		N	1992	2018	2123	1957	1851	2016	1913	1791
	Non-Hispanic Pacific Islander	Mean	307.02	317.7	288.81	259.11	283.4	303.22	314.54	297.09
		N	3377	3477	3714	3431	3161	3476	3201	3016
	Non-Hispanic White	Mean	332.07	328.74	295.97	278.79	305.63	312.29	329.77	310.06
		N	40797	41285	44045	40462	37886	41277	38680	36077
	Unknown	Mean	310.07	318.72	277.19	257.95	283.79	297.73	315.95	292.95
		N	26335	27082	28677	26167	24396	27067	25199	23451

Table 1.2.1.3

Mean Scale Scores by Ethnicity, S601 Online, continued

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
4–5	Hispanic (of any Race)	Mean	401.13	345.23	324.38	314.86	358.3	334.8	362.11	341.76
		N	239257	238232	243910	237435	224295	229577	226046	206794
	Non-Hispanic American Indian	Mean	403.8	342.3	317.51	311	358.55	330.33	361.5	339.7
		N	2574	2592	2641	2561	2380	2474	2420	2166
	Non-Hispanic Asian	Mean	418.35	359.79	341.47	326.68	372.85	350.68	377.55	357.21
		N	38181	38027	38795	37854	35926	36672	36243	33251
	Non-Hispanic Black	Mean	410.4	348.45	327.83	330.21	370.7	338.02	367.18	347.93
		N	18492	18369	19067	18335	17106	17686	17250	15576
	Non-Hispanic Multiracial	Mean	411.83	353.7	329.61	322.5	367.32	341.5	371.14	348.76
		N	1473	1484	1506	1473	1391	1434	1407	1296
	Non-Hispanic Pacific Islander	Mean	399.05	343.02	326.99	310.95	355.15	334.65	359.88	340.68
		N	3464	3436	3528	3473	3202	3241	3187	2853
	Non-Hispanic White	Mean	408.07	351.74	330.51	323.25	365.96	340.96	368.68	348.23
		N	32352	31870	32752	31986	29979	30427	30099	27061

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
	Unknown	Mean	387.9	338.92	309.62	296.63	342.15	323.7	353.54	328.73
		N	29037	29093	30514	28667	26716	28337	27179	24682
6–8	Hispanic (of any Race)	Mean	389.69	347.37	315.57	308.03	349.09	331.51	360.27	336.6
		N	281476	283727	291116	279581	260147	273231	265974	240854
	Non-Hispanic American Indian	Mean	394.93	347.2	316.82	307.07	351.31	332.16	361.93	337.99
		N	3358	3426	3494	3310	3080	3297	3205	2867
	Non-Hispanic Asian	Mean	402.61	362.22	328.91	327.62	365.4	345.6	374.6	351.49
		N	32743	32700	33503	32275	30106	31332	30704	27633
	Non-Hispanic Black	Mean	399.39	352.22	315.75	323.63	361.79	333.92	366.6	342.14
		N	20251	20379	20951	20123	18482	19437	18876	16770
	Non-Hispanic Multiracial	Mean	398.96	355.51	320.15	320.32	359.68	337.77	369.02	344.04
		N	1418	1437	1451	1394	1306	1376	1345	1205
	Non-Hispanic Pacific Islander	Mean	389.64	348.52	319.22	308.49	349.76	333.84	361.48	338.95
		N	3815	3882	4142	3951	3404	3646	3432	2989
	Non-Hispanic White	Mean	396.69	353.41	320.33	320.56	358.85	336.87	366.64	343.26
		N	32810	32839	33872	32481	29890	31269	30525	27046
	Unknown	Mean	380.58	342.04	303.42	294.83	337.44	322.4	353.64	326.29
		N	37919	38564	39749	37568	34478	36897	35529	31599
9–12	Hispanic (of any Race)	Mean	384.56	375.48	340.58	299.84	342.32	358.1	378.32	353.02
		N	299355	296259	306569	297256	280281	286569	280940	257899
	Non-Hispanic American Indian	Mean	396.23	380.97	346.97	307.11	352.02	363.93	385.61	359.72
		N	3600	3553	3699	3522	3312	3456	3358	3036
	Non-Hispanic Asian	Mean	405.94	393.05	358.12	329.38	367.77	375.65	397.07	372.89
		N	33393	32515	33511	32599	30941	31150	31065	28076
	Non-Hispanic Black	Mean	399.77	384.37	344.52	322.33	361.19	364.57	389.14	363.18

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
		N	25434	24791	26042	25149	23539	23890	23424	21272
	Non-Hispanic Multiracial	Mean	397.19	384.13	347.18	318.21	357.88	365.79	388.27	363.22
		N	1356	1350	1383	1332	1267	1308	1281	1186
	Non-Hispanic Pacific Islander	Mean	391.35	376.85	349.31	303.91	347.82	363.28	381.48	358.37
		N	3873	3736	3942	3759	3477	3539	3489	3066
	Non-Hispanic White	Mean	399.59	384.74	346.24	316.72	358.26	365.61	389.39	363.11
		N	35046	34382	35573	34212	32448	33185	32816	29773
	Unknown	Mean	379.99	372.23	330.36	290.95	335.33	351.15	374.55	345.67
		N	43758	43485	45527	43557	40386	41887	40523	36724

1.2.2 Mean Scale Score Across Domain and Composite Score by Grade

This section provides parallel information to the prior section, with mean scale scores broken down by grade rather than by grade-level cluster. Table 1.2.2.1 shows the increment of scale scores by grade, which peaked at Grade 5 in the Listening and Writing domains. The Clusters of 6–8 and 9–12 showed lower mean scale scores due to newcomers and long-term English learners (ELs).

Table 1.2.2.1

Mean Scale Scores by Grade, S601 Online

Grade		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Mean	303.26	283.68	237.99	240.09	271.81	260.9	289.56	263.97
	N	213228	218896	227850	210557	198705	218829	206299	192851
2	Mean	310.61	317.44	276.16	263.54	287.23	296.75	315.36	293.69
	N	207841	213322	224521	206974	193489	213228	199328	186234
	Mean	331.44	328.44	296.98	279.05	305.31	312.77	329.37	310.21

Grade		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
3	N	207218	210005	221384	206017	194511	209939	198479	187098
4	Mean	401	345.48	322	316.35	359.05	333.65	362.28	341.25
	N	199010	198383	203346	197760	186045	190734	187543	170696
5	Mean	405.42	348.77	330.06	315.88	360.85	339.38	365.84	345.62
	N	165820	164720	169367	164024	154950	159114	156288	142983
6	Mean	386.34	343.4	309.07	308.96	347.97	326.23	356.46	332.63
	N	140120	141417	145638	139307	128699	135779	131605	118160
7	Mean	391.98	349.38	317.01	309.94	351.25	333.18	362.38	338.46
	N	140171	141396	145443	138874	128759	135945	132021	118892
8	Mean	394.8	353.85	322.08	311.64	353.18	337.94	366.32	342.1
	N	133499	134141	137197	132502	123435	128761	125964	113911
9	Mean	382.03	372.01	336.33	296.84	339.55	354.13	375.1	349.39
	N	154657	153136	159048	153830	144063	147803	144409	131765
10	Mean	387.43	377.14	340.87	303.35	345.48	359.06	380.37	354.6
	N	125788	124032	128689	124507	117258	119687	117455	107284
11	Mean	393.41	383.25	347.14	310.91	352.25	365.28	386.4	360.88
	N	90901	89417	92588	88858	84028	86316	84990	77102
12	Mean	394.47	384.13	347.5	310.95	352.73	365.95	387.36	361.58
	N	74469	73486	75921	74191	70302	71178	70042	64881

Table 1.2.2.2

Mean Scale Scores by Grade by Gender, S601 Online

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	F	Mean	306.1	283.68	241.93	245.61	275.93	262.85	290.39	266.47
		N	88270	89959	93501	87395	83081	89935	85487	80693
	M	Mean	300.09	283.94	234.85	235.33	267.9	259.49	288.82	261.86
		N	95972	99000	102669	94755	89264	98960	93128	86845
	Missing	Mean	305.12	282.81	236.5	239.02	272.06	259.7	289.5	263.25
		N	28986	29937	31680	28407	26360	29934	27684	25313
2	F	Mean	312.5	318.25	281.53	269.47	291.1	299.82	316.48	296.91
		N	86640	88161	92861	86502	81370	88133	82972	78209
	M	Mean	308.83	316.96	271.94	258.95	284.17	294.43	314.54	291.22
		N	93444	96266	100834	93002	86984	96217	90027	84054

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
	Missing	Mean	310.67	316.57	273.84	260.4	285.3	295.12	314.67	291.81
		N	27757	28895	30826	27470	25135	28878	26329	23971
3	F	Mean	331.85	329.53	302.94	284.57	308.31	316.3	330.23	313.49
		N	83736	84293	88798	83391	79138	84277	80202	76113
	M	Mean	330.33	327.54	292.39	274.8	302.68	310.05	328.46	307.58
		N	94093	95773	100386	93568	88417	95731	90582	85403
	Missing	Mean	333.81	328.22	294.85	276.91	305.14	311.55	329.85	309.22
		N	29389	29939	32200	29058	26956	29931	27695	25582
4	F	Mean	400.68	346.04	328.53	320.43	360.88	337.2	362.53	344.17
		N	79426	78583	80643	78912	74546	75541	74665	68223
	M	Mean	401.09	345.19	318.03	313.74	357.82	331.55	362.13	339.44
		N	91328	91466	93350	90813	85545	88079	86599	78960
	Missing	Mean	401.57	344.83	316.68	313.35	357.87	330.6	362.09	338.87
		N	28256	28334	29353	28035	25954	27114	26279	23513
5	F	Mean	404.27	349.47	336.31	317.68	361.15	342.83	365.94	348.02
		N	64517	63747	65503	63674	60490	61645	60833	55845
	M	Mean	405.23	347.66	325.67	314.5	360.08	336.62	365	343.46
		N	76825	76817	78540	76134	71907	74205	72820	66662
	Missing	Mean	409.04	350.49	327.49	315.47	362.51	339.05	368.31	346.14
		N	24478	24156	25324	24216	22553	23264	22635	20476

Table 1.2.2.2

Mean Scale Scores by Grade by Gender, S601 Online, continued

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
6	F	Mean	384.29	344.15	313.06	309.66	347.2	328.62	356.33	333.93
		N	54021	54063	55571	53410	49775	51959	50802	45772
	M	Mean	387.53	342.7	306.26	308.97	348.66	324.46	356.34	331.67
		N	65469	66392	68227	65410	60335	63787	61651	55474
	Missing	Mean	387.96	343.67	307.73	307.08	347.78	325.66	357.18	332.23
		N	20630	20962	21840	20487	18589	20033	19152	16914
7	F	Mean	390.34	350.88	321.05	311.26	351.03	335.95	362.87	340.16
		N	54000	53883	55437	53116	49678	51832	50771	45792
	M	Mean	393.28	348.53	314.43	309.75	351.87	331.47	362.2	337.53
		N	66099	67081	68769	65877	60964	64490	62508	56411
	Missing	Mean	392.11	348.21	314.84	307.04	349.76	331.52	361.68	336.96
		N	20072	20432	21237	19881	18117	19623	18742	16689
8	F	Mean	393.43	355.81	325.67	313.71	353.37	340.76	367.2	343.94
		N	50997	50926	51967	50256	47157	48882	48155	43558

	M	Mean	395.79	352.59	319.81	311.03	353.44	336.13	365.76	341.03
		N	62830	63362	64773	62713	58300	60868	59418	53865
	Missing	Mean	395.19	352.87	320.13	308.29	351.81	336.47	365.83	340.71
		N	19672	19853	20457	19533	17978	19011	18391	16488
9	F	Mean	381.21	374.12	340.06	299.93	340.61	357.1	376.35	351.73
		N	57648	56832	59001	57263	53864	54895	53832	49288
	M	Mean	381.83	369.96	333.83	294.37	338.28	351.84	373.6	347.44
		N	73853	73431	76137	73660	68809	70854	69113	63031
	Missing	Mean	384.69	373.3	335.07	297.1	340.99	354.12	376.78	349.75
		N	23156	22873	23910	22907	21390	22054	21464	19446
10	F	Mean	387	379.47	344.14	306.45	346.74	361.86	381.81	356.85
		N	46993	45944	47694	46313	43905	44366	43845	40184
	M	Mean	387.34	375.36	338.82	301.34	344.48	357.17	379.12	353.05
		N	60894	60492	62543	60590	56855	58374	57080	52119
	Missing	Mean	388.9	377.15	339.37	302.12	345.58	358.25	380.83	353.99
		N	17901	17596	18452	17604	16498	16947	16530	14981
11	F	Mean	393.05	385.32	350.35	313.99	353.6	367.96	387.78	363.12
		N	34940	34224	35309	34005	32354	33020	32735	29721
	M	Mean	393.83	381.71	345.14	309.15	351.6	363.51	385.46	359.53
		N	42803	42238	43722	42062	39744	40830	40090	36547
	Missing	Mean	392.97	382.81	345.24	308.52	350.77	364.01	385.82	359.3
		N	13158	12955	13557	12791	11930	12466	12165	10834
12	F	Mean	393.8	386.17	350.67	313.89	353.89	368.58	388.62	363.72
		N	29293	28780	29648	29095	27710	27861	27558	25544
	M	Mean	394.63	382.09	345.2	308.68	351.7	363.78	385.98	359.83
		N	34801	34450	35549	34789	32971	33415	32853	30546
	Missing	Mean	395.81	385.25	346.37	310.27	352.96	365.87	388.47	361.44
		N	10375	10256	10724	10307	9621	9902	9631	8791

Table 1.2.2.3

Mean Scale Scores by Grade by Ethnicity, S601 Online

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Hispanic (of any Race)	Mean	299.31	279.91	231.54	235.77	267.68	255.8	285.72	259.17
		N	137312	141152	146488	135707	128221	141101	133176	124716
	Non-Hispanic American Indian	Mean	307.6	282.27	232.09	241.77	274.82	257.52	290.25	263.11
		N	1355	1400	1464	1337	1253	1399	1298	1201
	Non-Hispanic Asian	Mean	318.7	299.19	263.12	254.17	286.72	281.36	305.13	282.95
		N	26672	27231	28406	26300	24868	27228	25776	24113
	Non-Hispanic Black	Mean	307.61	288.16	245.33	256.1	281.9	266.84	294.1	271.05
		N	10534	10822	11363	10350	9682	10815	10095	9305

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
	Non-Hispanic Multiracial	Mean	315.03	290.03	249.26	250.36	282.73	269.65	297.61	273.31
		N	1038	1071	1103	1028	974	1071	1015	957
	Non-Hispanic Pacific Islander	Mean	291.11	279.7	236.11	233.28	262.36	258	283.21	259.55
		N	1662	1729	1814	1681	1557	1728	1599	1511
	Non-Hispanic White	Mean	313.95	288.66	250.2	248.6	281.51	269.48	296.24	272.93
		N	20857	21254	22321	20492	19314	21250	20016	18611
	Unknown	Mean	293.36	280.76	229.53	230.97	262.02	254.99	284.45	256.5
		N	13798	14237	14891	13662	12836	14237	13324	12437
2	Hispanic (of any Race)	Mean	305.56	314.49	270.99	260.14	282.98	292.71	311.79	289.56
		N	133815	137678	144426	133453	124769	137617	128663	120340
	Non-Hispanic American Indian	Mean	314.37	315.02	272.67	263.53	289.12	294.05	314.82	292.55
		N	1383	1409	1482	1369	1291	1407	1318	1230
	Non-Hispanic Asian	Mean	329.83	329.68	296.93	275.86	303.15	313.4	329.78	310.28
		N	25565	26110	27432	25428	23915	26101	24583	23093
	Non-Hispanic Black	Mean	317.7	320.36	281.54	277.56	298.07	300.98	319.63	300.09
		N	10707	11026	11659	10679	9915	11021	10222	9505
	Non-Hispanic Multiracial	Mean	328.01	324.95	287.33	274.59	301.57	305.97	325.96	304.27
		N	989	1002	1058	971	917	1001	947	888
	Non-Hispanic Pacific Islander	Mean	297.53	313	277.77	252.38	275.17	295.19	308.37	288.84
		N	1599	1651	1762	1624	1491	1650	1516	1421
	Non-Hispanic White	Mean	322.18	322.03	286.63	271.01	296.8	304.19	322.01	301.75
		N	20455	20725	22163	20290	18923	20720	19332	17975
	Unknown	Mean	300.79	314.67	267.51	251.83	276.08	290.79	310.29	285.81
		N	13328	13721	14539	13160	12268	13711	12747	11782

Table 1.2.2.3

Mean Scale Scores by Grade by Ethnicity, S601 Online, continued

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
3	Hispanic (of any Race)	Mean	327.13	324.75	293.09	276.32	301.79	308.98	325.49	306.51
		N	133606	135171	142221	132861	125765	135123	128099	121064
	Non-Hispanic American	Mean	332.12	322.78	291.81	273.43	302.64	307.46	325.81	305.81
		N	1503	1573	1637	1526	1409	1571	1456	1367
	Non-Hispanic Asian	Mean	349.94	344.22	314.37	290.7	320.5	329.52	346.04	326.62
		N	25167	25496	26808	24963	23641	25494	24163	22798

	Non-Hispanic Black	Mean	337.43	331.73	300.34	291.76	314.73	316.13	333.41	315.44
		N	10812	11002	11681	10695	10001	10997	10310	9600
	Non-Hispanic Multiracial	Mean	344.27	337.37	304.77	290.73	317.89	321.27	339.52	320.39
		N	1003	1016	1065	986	934	1015	966	903
	Non-Hispanic Pacific Islander	Mean	315.56	321.95	298.77	265.17	290.75	310.47	320.09	304.43
		N	1778	1826	1952	1807	1670	1826	1685	1595
	Non-Hispanic White	Mean	342.01	335.5	305.42	286.6	314.44	320.45	337.52	318.31
		N	20342	20560	21882	20172	18963	20557	19348	18102
	Unknown	Mean	319.57	322.89	287.15	264.15	291.6	304.86	321.75	300.17
		N	13007	13361	14138	13007	12128	13356	12452	11669
4	Hispanic (of any Race)	Mean	398.52	343.3	320.06	314.65	356.94	331.66	360.01	339.22
		N	128766	128451	131401	128108	120693	123536	121593	110989
	Non-Hispanic American	Mean	399.09	338.9	309.35	308.73	355.16	324.64	357.88	334.6
		N	1343	1368	1389	1354	1239	1299	1258	1122
	Non-Hispanic Asian	Mean	417.21	358.68	339.01	327.08	372.51	348.86	376.41	355.87
		N	22461	22374	22780	22299	21152	21525	21313	19522
	Non-Hispanic Black	Mean	408.33	347.36	324.35	330.17	369.79	335.71	365.74	346.06
		N	10250	10210	10561	10170	9467	9802	9571	8597
	Non-Hispanic Multiracial	Mean	410.16	352.91	327	323.23	366.75	339.86	370.18	347.43
		N	850	850	857	838	798	820	812	744
	Non-Hispanic Pacific Islander	Mean	394.46	340.55	321.43	307.59	351.35	330.82	357.02	336.96
		N	1841	1807	1842	1812	1708	1709	1707	1534
	Non-Hispanic White	Mean	406.67	350.73	327.95	323.99	365.76	339.13	367.61	346.95
		N	18144	17897	18342	18000	16857	17028	16896	15153
	Unknown	Mean	386.9	337.9	306.43	298.02	342.63	321.47	352.58	327.55
		N	15355	15426	16174	15179	14131	15015	14393	13035

Table 1.2.2.3

Mean Scale Scores by Grade by Ethnicity, S601 Online, continued

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
5	Hispanic (of any Race)	Mean	404.17	347.48	329.42	315.11	359.89	338.46	364.55	344.7
		N	110491	109781	112509	109327	103602	106041	104453	95805
	Non-Hispanic American Indian	Mean	408.95	346.09	326.57	313.54	362.23	336.61	365.41	345.19
		N	1231	1224	1252	1207	1141	1175	1162	1044

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
	Non-Hispanic Asian	Mean	419.98	361.37	344.97	326.1	373.34	353.27	379.18	359.11
		N	15720	15653	16015	15555	14774	15147	14930	13729
	Non-Hispanic Black	Mean	412.97	349.81	332.16	330.27	371.84	340.9	368.97	350.23
		N	8242	8159	8506	8165	7639	7884	7679	6979
	Non-Hispanic Multiracial	Mean	414.11	354.76	333.05	321.53	368.09	343.68	372.46	350.54
		N	623	634	649	635	593	614	595	552
	Non-Hispanic Pacific Islander	Mean	404.27	345.77	333.07	314.62	359.5	338.92	363.18	345.02
		N	1623	1629	1686	1661	1494	1532	1480	1319
	Non-Hispanic White	Mean	409.86	353.04	333.77	322.29	366.22	343.29	370.06	349.87
		N	14208	13973	14410	13986	13122	13399	13203	11908
	Unknown	Mean	389.02	340.07	313.21	295.06	341.6	326.22	354.62	330.06
		N	13682	13667	14340	13488	12585	13322	12786	11647
6	Hispanic (of any Race)	Mean	385.54	342.19	308.9	307.69	346.94	325.58	355.37	331.86
		N	95137	95966	98686	94554	87670	92342	89677	80910
	Non-Hispanic American Indian	Mean	389.13	341.61	309.85	307.71	349.14	325.84	356.14	333.06
		N	1093	1115	1124	1082	1005	1065	1041	930
	Non-Hispanic Asian	Mean	395.45	354.69	320.47	321.64	358.95	337.65	367.12	344
		N	11382	11427	11750	11255	10453	10973	10678	9627
	Non-Hispanic Black	Mean	392.38	345.31	306.96	319.72	356.45	326.07	359.69	335.11
		N	6683	6826	7018	6723	6080	6507	6246	5518
	Non-Hispanic Multiracial	Mean	391.8	349.26	311.91	314.69	353.74	330.86	362.96	337.56
		N	479	491	499	474	439	470	451	403
	Non-Hispanic Pacific Islander	Mean	384.4	343.8	312.84	306.37	346.28	328.22	356.36	334.02
		N	1356	1377	1477	1420	1224	1304	1211	1064
	Non-Hispanic White	Mean	390.71	347.25	313.14	317.66	354.51	330.11	360.46	337.36
		N	11261	11249	11692	11176	10265	10724	10423	9216
	Unknown	Mean	376.91	337.96	297.31	293.91	335.47	317.32	349.66	322.26
		N	12729	12966	13392	12623	11563	12394	11878	10492

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
7	Hispanic (of any Race)	Mean	390.31	347.67	316.39	307.26	349.11	332.08	360.68	337.05
		N	95027	95923	98533	94214	87654	92500	89845	81349
	Non-Hispanic American Indian	Mean	396.2	347.15	318.02	306.34	351.41	332.8	362.44	338.52
		N	1076	1113	1132	1058	981	1073	1030	919
	Non-Hispanic Asian	Mean	404.75	364.11	330.85	328.91	367.12	347.47	376.61	353.42
		N	11353	11344	11642	11197	10434	10870	10655	9581
	Non-Hispanic Black	Mean	400.88	352.85	317.2	324.35	363.03	334.86	367.46	343.33
		N	6887	6873	7106	6791	6279	6566	6394	5691
	Non-Hispanic Multiracial	Mean	400.33	357	321.63	322.91	361.6	339.02	369.95	345.42
		N	492	500	499	481	450	475	468	414
	Non-Hispanic Pacific Islander	Mean	389.7	347.82	319.5	308.39	349.65	333.61	361.16	338.74
		N	1289	1319	1414	1342	1151	1227	1157	998
	Non-Hispanic White	Mean	398.46	354.81	322.26	321.28	360.2	338.53	368.18	344.77
		N	11158	11202	11584	11059	10138	10675	10377	9193
	Unknown	Mean	382.18	342.7	304.51	295.53	338.74	323.28	354.71	327.51
		N	12889	13122	13533	12732	11672	12559	12095	10747
8	Hispanic (of any Race)	Mean	393.36	352.46	321.73	309.19	351.29	337.1	364.93	341
		N	91312	91838	93897	90813	84823	88389	86452	78595
	Non-Hispanic American Indian	Mean	399.11	352.46	322.06	307.15	353.22	337.38	366.78	342
		N	1189	1198	1238	1170	1094	1159	1134	1018
	Non-Hispanic Asian	Mean	408.31	368.72	336.46	332.99	370.77	352.65	380.84	357.86
		N	10008	9929	10111	9823	9219	9489	9371	8425
	Non-Hispanic Black	Mean	404.88	358.61	323.29	326.86	365.8	340.96	372.63	347.9
		N	6681	6680	6827	6609	6123	6364	6236	5561
	Non-Hispanic Multiracial	Mean	405.12	360.7	327.61	323.56	363.86	343.93	374.42	349.29
		N	447	446	453	439	417	431	426	388
	Non-Hispanic Pacific Islander	Mean	395.64	354.77	326.44	311.11	354.03	340.65	367.64	344.83
		N	1170	1186	1251	1189	1029	1115	1064	927
	Non-Hispanic White	Mean	401.26	358.58	326.15	322.94	362.09	342.4	371.61	347.96
		N	10391	10388	10596	10246	9487	9870	9725	8637
	Unknown	Mean	382.69	345.61	308.64	295.04	338.12	326.77	356.62	329.09
		N	12301	12476	12824	12213	11243	11944	11556	10360

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
9	Hispanic (of any Race)	Mean	379.51	370.27	335.68	293.89	336.88	352.96	373.14	347.84
		N	106198	105384	109231	105886	99304	101854	99470	91125
	Non-Hispanic American Indian	Mean	391.96	374.78	343.8	301.27	347.19	359.07	379.97	354.89
		N	1238	1237	1284	1220	1137	1208	1164	1055
	Non-Hispanic Asian	Mean	400.75	387.81	353.57	323.06	362	370.63	391.84	367.55
		N	10419	10193	10524	10204	9609	9756	9703	8722
	Non-Hispanic Black	Mean	394.43	379.09	340.61	316.38	355.48	359.86	383.78	358.04
		N	7722	7534	7957	7660	7131	7253	7101	6417
	Non-Hispanic Multiracial	Mean	391.81	377.78	342.98	311.64	352.15	360.38	382.28	357.78
		N	495	497	507	487	463	485	470	439
	Non-Hispanic Pacific Islander	Mean	388.02	373.64	348.71	299.54	343.5	361.38	378.15	355.59
		N	1406	1375	1432	1380	1266	1295	1273	1112
	Non-Hispanic White	Mean	392.92	378.67	342.29	310.5	351.74	360.61	383.01	357.64
		N	11565	11347	11801	11362	10703	10950	10782	9763
	Unknown	Mean	370.82	364.59	321.31	279.22	324.73	342.66	366.26	336.46
		N	15614	15569	16312	15631	14450	15002	14446	13132
10	Hispanic (of any Race)	Mean	384	374.87	339.65	299.23	341.71	357.34	377.74	352.29
		N	85982	85014	87995	85382	80499	82150	80578	73920
	Non-Hispanic American Indian	Mean	397.49	382.22	345.79	306.76	352.52	363.71	386.99	359.85
		N	1043	1027	1068	1025	956	984	960	854
	Non-Hispanic Asian	Mean	405.86	392.78	357.04	328.95	367.46	375.04	396.84	372.42
		N	9163	8893	9211	8938	8478	8504	8482	7637
	Non-Hispanic Black	Mean	400.67	384.41	344.18	322.57	361.82	364.47	389.51	363.37
		N	6723	6585	6871	6619	6195	6331	6204	5606
	Non-Hispanic Multiracial	Mean	394.95	382.87	343.97	314.95	354.08	363.57	385.95	359.78
		N	390	386	396	377	360	372	367	336
	Non-Hispanic Pacific Islander	Mean	390.22	375.73	348.95	302.97	346.95	362.88	380.34	357.79
		N	1013	960	1034	964	891	905	889	770
	Non-Hispanic White	Mean	398.8	383.81	345.57	316.89	357.93	364.78	388.51	362.48
		N	9650	9425	9772	9412	8952	9100	9018	8196
	Unknown	Mean	379.98	371.73	330.71	291.61	335.75	351.1	374.23	345.86
		N	11824	11742	12342	11790	10927	11341	10957	9965

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
11	Hispanic (of any Race)	Mean	389.59	380.71	345.98	306.25	347.99	363.44	383.45	358.33
		N	59616	58786	60759	58500	55434	56872	55982	51088
	Non-Hispanic American Indian	Mean	400.28	387.98	351.01	315.2	357.68	369.54	391.9	365.14
		N	718	692	734	687	655	679	662	597
	Non-Hispanic Asian	Mean	410.03	397.7	363.2	335.86	372.98	380.55	401.44	377.84
		N	7252	7066	7241	6998	6664	6767	6757	6064
	Non-Hispanic Black	Mean	404.28	389.13	348.33	328.69	366.77	368.93	393.89	367.92
		N	5710	5516	5813	5568	5235	5315	5227	4718
	Non-Hispanic Multiracial	Mean	405.67	392.46	354.81	330.32	368.69	374.03	397.13	372.88
		N	284	278	287	276	261	267	264	239
	Non-Hispanic Pacific Islander	Mean	394.1	379.98	350.57	307.59	351.17	365.02	383.97	360.11
		N	767	742	777	734	685	709	703	618
	Non-Hispanic White	Mean	404.72	389.82	350.66	322.44	363.87	370.41	394.66	368.12
		N	7430	7285	7512	7155	6804	7008	6951	6243
	Unknown	Mean	388.11	379.2	337.98	300.9	344.5	358.53	381.97	353.5
		N	9124	9052	9465	8940	8290	8699	8444	7535
12	Hispanic (of any Race)	Mean	390.52	381.68	346.54	306.32	348.44	364.28	384.45	359.11
		N	47559	47075	48584	47488	45044	45693	44910	41766
	Non-Hispanic American Indian	Mean	398.01	383.54	350.83	310.41	354.3	367.82	387.5	362.99
		N	601	597	613	590	564	585	572	530
	Non-Hispanic Asian	Mean	409.78	396.67	361.35	332.95	371.54	379.08	400.87	376.47
		N	6559	6363	6535	6459	6190	6123	6123	5653
	Non-Hispanic Black	Mean	401.58	386.91	346.61	323.93	362.73	366.89	391.35	365.31
		N	5279	5156	5401	5302	4978	4991	4892	4531
	Non-Hispanic Multiracial	Mean	403.22	391.13	353.5	323.9	364.43	372.57	395.65	370.35
		N	187	189	193	192	183	184	180	172
	Non-Hispanic Pacific Islander	Mean	396.77	381.63	349.67	310.11	354.07	365.8	387.07	362.69
		N	687	659	699	681	635	630	624	566
	Non-Hispanic White	Mean	406.88	391.16	349.35	321.21	364.05	370.3	395.99	368.02
		N	6401	6325	6488	6283	5989	6127	6065	5571
	Unknown	Mean	389.61	380.9	339.96	302.96	346.15	360.45	383.61	355.53
		N	7196	7122	7408	7196	6719	6845	6676	6092

1.2.3 Correlations

Tables in this section show Pearson correlations among the four domain scale scores by grade-level cluster across all tiers, as well as the number of students included in each correlation. The pattern of domain correlations varied across clusters. In Grade 1, Listening was correlated to Speaking; Reading was correlated to Writing. In Clusters 2–3, Listening was mostly correlated to Speaking and Writing, and Reading was correlated to Listening. In Clusters 4–5 and 6–8, Listening was correlated to Reading and Writing and Reading was correlated to Listening and Writing. In Cluster 9–12, the Listening and Reading domains were highly correlated and the Listening, Reading, and Writing domains were correlated to the Speaking domain.

Table 1.2.3.1

Correlations Among Scale Scores: Grade 1, S601 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Speaking	Writing
Listening	Pearson Correlation	1	0.381	0.596	0.529
	N	213228	206299	198705	213162
Reading	Pearson Correlation		1	0.34	0.472
	N		218896	203348	218829
Speaking	Pearson Correlation			1	0.474
	N			210557	210488
Writing	Pearson Correlation				1
	N				227850

Table 1.2.3.2

Correlations Among Scale Scores: Grades 2–3, S601 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Speaking	Writing
Listening	Pearson Correlation	1	0.579	0.623	0.608
	N	415059	397807	388000	414895
Reading	Pearson Correlation		1	0.466	0.551
	N		423327	394797	423167
Speaking	Pearson Correlation			1	0.561
	N			412991	412840
Writing	Pearson Correlation				1
	N				445905

Table 1.2.3.3

Correlations Among Scale Scores: Grades 4–5, S601 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Speaking	Writing
Listening	Pearson Correlation	1	0.664	0.625	0.668
	N	364830	343831	340995	350259
Reading	Pearson Correlation		1	0.511	0.664
	N		363103	339613	349848
Speaking	Pearson Correlation			1	0.626
	N			361784	347170
Writing	Pearson Correlation				1
	N				372713

Table 1.2.3.4

Correlations Among Scale Scores: Grades 6–8, S601 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Speaking	Writing
Listening	Pearson Correlation	1	0.669	0.605	0.633
	N	413790	389590	380893	396150
Reading	Pearson Correlation		1	0.537	0.643
	N		416954	384077	400485
Speaking	Pearson Correlation			1	0.631
	N			410683	392580
Writing	Pearson Correlation				1
	N				428278

Table 1.2.3.5

Correlations Among Scale Scores: Grades 9–12, S601 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Speaking	Writing
Listening	Pearson Correlation	1	0.716	0.607	0.57
	N	445815	416896	415651	428781
Reading	Pearson Correlation		1	0.604	0.586
	N		440071	411784	424984
Speaking	Pearson Correlation			1	0.637
	N			441386	424437
Writing	Pearson Correlation				1
	N				456246

1.3 Proficiency Level Results

The performance by domain was observed in the descending order of Listening, Reading, Speaking, and Writing. For Listening, there was a large percentage (66%) in Proficiency Level (PL) 6, especially in Cluster 4–5. The Reading domain had 4% to 10% in PL 6. For the Writing domain, fewer than 1% of students were in PL 5 and PL 6 together, except Cluster 4–5 showed 3% in PL 5 and 6. In the Speaking domain, fewer than 1% were in PL 5 and PL 6; Cluster 4–5 showed 2% in both PL ranges.

1.3.1 Domains

1.3.1.1 Listening

1.3.1.1.1 By Cluster

Table 1.3.1.1.1.1

Proficiency Level by Cluster (Count): Listening, S601 Online

	Listening Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	33262	15847	32747	14135	29086	88151	213228
2–3	62746	48216	83242	44488	72990	103377	415059
4–5	12677	16773	29093	19270	45440	241577	364830
6–8	17902	44411	68782	69660	89229	123806	413790
9–12	50105	67833	103482	92025	69967	62403	445815

Table 1.3.1.1.1.2

Proficiency Level by Cluster (Percent): Listening, S601 Online

	Listening Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	15.60%	7.40%	15.40%	6.60%	13.60%	41.30%	100.00%
2–3	15.10%	11.60%	20.10%	10.70%	17.60%	24.90%	100.00%
4–5	3.50%	4.60%	8.00%	5.30%	12.50%	66.20%	100.00%
6–8	4.30%	10.70%	16.60%	16.80%	21.60%	29.90%	100.00%
9–12	11.20%	15.20%	23.20%	20.60%	15.70%	14.00%	100.00%

1.3.1.1.2 By Grade

Table 1.3.1.1.2.1

Proficiency Level by Grade (Count): Listening, S601 Online

Grade	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	33262	15847	32747	14135	29086	88151	213228
2	32690	24900	41251	21771	37948	49281	207841
3	30056	23316	41991	22717	35042	54096	207218
4	4606	8044	15851	9413	22518	138578	199010
5	8071	8729	13242	9857	22922	102999	165820
6	3563	13721	22759	23019	33957	43101	140120
7	5803	15072	23596	23588	28895	43217	140171
8	8536	15618	22427	23053	26377	37488	133499
9	13531	27227	35305	30679	24533	23382	154657
10	13266	18670	31437	25180	19440	17795	125788
11	12281	11659	20125	19535	15621	11680	90901
12	11027	10277	16615	16631	10373	9546	74469

Table 1.3.1.1.2.2

Proficiency Level by Grade (Percent): Listening, S601 Online

Grade	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	15.60%	7.40%	15.40%	6.60%	13.60%	41.30%	100.00%
2	15.70%	12.00%	19.80%	10.50%	18.30%	23.70%	100.00%
3	14.50%	11.30%	20.30%	11.00%	16.90%	26.10%	100.00%
4	2.30%	4.00%	8.00%	4.70%	11.30%	69.60%	100.00%
5	4.90%	5.30%	8.00%	5.90%	13.80%	62.10%	100.00%
6	2.50%	9.80%	16.20%	16.40%	24.20%	30.80%	100.00%
7	4.10%	10.80%	16.80%	16.80%	20.60%	30.80%	100.00%
8	6.40%	11.70%	16.80%	17.30%	19.80%	28.10%	100.00%
9	8.70%	17.60%	22.80%	19.80%	15.90%	15.10%	100.00%
10	10.50%	14.80%	25.00%	20.00%	15.50%	14.10%	100.00%
11	13.50%	12.80%	22.10%	21.50%	17.20%	12.80%	100.00%
12	14.80%	13.80%	22.30%	22.30%	13.90%	12.80%	100.00%

1.3.1.2 Reading

1.3.1.2.1 By Cluster

Table 1.3.1.2.1.1

Proficiency Level by Cluster (Count): Reading, S601 Online

	Reading Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	58284	73543	40204	19911	13386	13568	218896
2–3	58534	120630	80905	57340	63968	41950	423327
4–5	55204	82233	79493	52431	55775	37967	363103
6–8	134175	108252	94603	28999	33144	17781	416954
9–12	93868	116492	94383	37236	53507	44585	440071

Table 1.3.1.2.1.2

Proficiency Level by Cluster (Percent): Reading, S601 Online

	Reading Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	26.60%	33.60%	18.40%	9.10%	6.10%	6.20%	100.00%
2–3	13.80%	28.50%	19.10%	13.50%	15.10%	9.90%	100.00%
4–5	15.20%	22.60%	21.90%	14.40%	15.40%	10.50%	100.00%
6–8	32.20%	26.00%	22.70%	7.00%	7.90%	4.30%	100.00%
9–12	21.30%	26.50%	21.40%	8.50%	12.20%	10.10%	100.00%

1.3.1.2.2 By Grade

Table 1.3.1.2.2.1

Proficiency Level by Grade (Count): Reading, S601 Online

	Reading Proficiency Range						
Grade	1	2	3	4	5	6	Total
1	58284	73543	40204	19911	13386	13568	218896
2	19919	59654	50092	33017	34105	16535	213322
3	38615	60976	30813	24323	29863	25415	210005
4	25190	44423	39960	33784	32267	22759	198383
5	30014	37810	39533	18647	23508	15208	164720
6	42595	39111	35403	9894	10536	3878	141417
7	45116	37383	31546	9078	11998	6275	141396
8	46464	31758	27654	10027	10610	7628	134141
9	32298	41625	32014	12074	19329	15796	153136

Grade	Reading Proficiency Range						Total
	1	2	3	4	5	6	
10	26278	33177	25454	12091	14895	12137	124032
11	18237	22592	19962	7512	11252	9862	89417
12	17055	19098	16953	5559	8031	6790	73486

Table 1.3.1.2.2.2

Proficiency Level by Grade (Percent): Reading, S601 Online

Grade	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	26.60%	33.60%	18.40%	9.10%	6.10%	6.20%	100.00%
2	9.30%	28.00%	23.50%	15.50%	16.00%	7.80%	100.00%
3	18.40%	29.00%	14.70%	11.60%	14.20%	12.10%	100.00%
4	12.70%	22.40%	20.10%	17.00%	16.30%	11.50%	100.00%
5	18.20%	23.00%	24.00%	11.30%	14.30%	9.20%	100.00%
6	30.10%	27.70%	25.00%	7.00%	7.50%	2.70%	100.00%
7	31.90%	26.40%	22.30%	6.40%	8.50%	4.40%	100.00%
8	34.60%	23.70%	20.60%	7.50%	7.90%	5.70%	100.00%
9	21.10%	27.20%	20.90%	7.90%	12.60%	10.30%	100.00%
10	21.20%	26.70%	20.50%	9.70%	12.00%	9.80%	100.00%
11	20.40%	25.30%	22.30%	8.40%	12.60%	11.00%	100.00%
12	23.20%	26.00%	23.10%	7.60%	10.90%	9.20%	100.00%

1.3.1.3 Writing

1.3.1.3.1 By Cluster

Table 1.3.1.3.1.1

Proficiency Level by Cluster (Count): Writing, S601 Online

Cluster	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	102504	77067	45839	2427	8	5	227850
2–3	70848	100727	237739	36072	508	11	445905
4–5	37968	26641	193901	105621	7511	1071	372713
6–8	63268	83904	244120	36874	110	2	428278
9–12	62191	95545	234822	62508	1149	31	456246

Table 1.3.1.3.1.2

Proficiency Level by Cluster (Percent): Writing, S601 Online

Cluster	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	45.00%	33.80%	20.10%	1.10%	0.00%	0.00%	100.00%
2–3	15.90%	22.60%	53.30%	8.10%	0.10%	0.00%	100.00%
4–5	10.20%	7.10%	52.00%	28.30%	2.00%	0.30%	100.00%
6–8	14.80%	19.60%	57.00%	8.60%	0.00%	0.00%	100.00%
9–12	13.60%	20.90%	51.50%	13.70%	0.30%	0.00%	100.00%

1.3.1.3.2 By Grade

Table 1.3.1.3.2.1

Proficiency Level by Grade (Count): Writing, S601 Online

Grade	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	102504	77067	45839	2427	8	5	227850
2	41876	64027	109627	8935	56	0	224521
3	28972	36700	128112	27137	452	11	221384
4	20275	15688	97661	63806	5527	389	203346
5	17693	10953	96240	41815	1984	682	169367
6	18887	36566	75347	14822	16	0	145638
7	23311	23212	89968	8882	69	1	145443
8	21070	24126	78805	13170	25	1	137197
9	20038	29073	79272	30139	502	24	159048
10	14976	26534	74912	11829	436	2	128689
11	11767	25335	43997	11348	138	3	92588
12	15410	14603	36641	9192	73	2	75921

Table 1.3.1.3.2.2

Proficiency Level by Grade (Percent): Writing, S601 Online

Grade	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	45.00%	33.80%	20.10%	1.10%	0.00%	0.00%	100.00%
2	18.70%	28.50%	48.80%	4.00%	0.00%	0.00%	100.00%

3	13.10%	16.60%	57.90%	12.30%	0.20%	0.00%	100.00%
4	10.00%	7.70%	48.00%	31.40%	2.70%	0.20%	100.00%
5	10.40%	6.50%	56.80%	24.70%	1.20%	0.40%	100.00%
6	13.00%	25.10%	51.70%	10.20%	0.00%	0.00%	100.00%
7	16.00%	16.00%	61.90%	6.10%	0.00%	0.00%	100.00%
8	15.40%	17.60%	57.40%	9.60%	0.00%	0.00%	100.00%
9	12.60%	18.30%	49.80%	18.90%	0.30%	0.00%	100.00%
10	11.60%	20.60%	58.20%	9.20%	0.30%	0.00%	100.00%
11	12.70%	27.40%	47.50%	12.30%	0.10%	0.00%	100.00%
12	20.30%	19.20%	48.30%	12.10%	0.10%	0.00%	100.00%

1.3.1.4 Speaking

1.3.1.4.1 By Cluster

Table 1.3.1.4.1.1

Proficiency Level by Cluster (Count): Speaking, S601 Online

	Speaking Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	48263	81992	58177	21213	831	81	210557
2–3	79141	121558	145167	64141	2432	552	412991
4–5	45047	68526	124461	108351	14178	1221	361784
6–8	98843	101484	150440	58112	1681	123	410683
9–12	177984	96474	146338	19749	604	237	441386

Table 1.3.1.4.1.2

Proficiency Level by Cluster (Percent): Speaking, S601 Online

	Speaking Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	22.90%	38.90%	27.60%	10.10%	0.40%	0.00%	100.00%
2–3	19.20%	29.40%	35.20%	15.50%	0.60%	0.10%	100.00%
4–5	12.50%	18.90%	34.40%	29.90%	3.90%	0.30%	100.00%
6–8	24.10%	24.70%	36.60%	14.20%	0.40%	0.00%	100.00%
9–12	40.30%	21.90%	33.20%	4.50%	0.10%	0.10%	100.00%

1.3.1.4.2 By Grade

Table 1.3.1.4.2.1

Proficiency Level by Grade (Count): Speaking, S601 Online

Grade	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	48263	81992	58177	21213	831	81	210557
2	37929	72058	67848	27666	1348	125	206974
3	41212	49500	77319	36475	1084	427	206017
4	19398	35483	68087	64619	9210	963	197760
5	25649	33043	56374	43732	4968	258	164024
6	29142	37576	50757	20924	894	14	139307
7	34406	29956	57055	16879	533	45	138874
8	35295	33952	42628	20309	254	64	132502
9	68284	30137	47420	7775	156	58	153830
10	49659	24233	45352	5065	140	58	124507
11	33438	21066	29305	4796	182	71	88858
12	26603	21038	24261	2113	126	50	74191

Table 1.3.1.4.2.2

Proficiency Level by Grade (Percent): Speaking, S601 Online

Grade	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	22.90%	38.90%	27.60%	10.10%	0.40%	0.00%	100.00%
2	18.30%	34.80%	32.80%	13.40%	0.70%	0.10%	100.00%
3	20.00%	24.00%	37.50%	17.70%	0.50%	0.20%	100.00%
4	9.80%	17.90%	34.40%	32.70%	4.70%	0.50%	100.00%
5	15.60%	20.10%	34.40%	26.70%	3.00%	0.20%	100.00%
6	20.90%	27.00%	36.40%	15.00%	0.60%	0.00%	100.00%
7	24.80%	21.60%	41.10%	12.20%	0.40%	0.00%	100.00%
8	26.60%	25.60%	32.20%	15.30%	0.20%	0.00%	100.00%
9	44.40%	19.60%	30.80%	5.10%	0.10%	0.00%	100.00%
10	39.90%	19.50%	36.40%	4.10%	0.10%	0.00%	100.00%
11	37.60%	23.70%	33.00%	5.40%	0.20%	0.10%	100.00%
12	35.90%	28.40%	32.70%	2.80%	0.20%	0.10%	100.00%

1.3.2 Composites

The observed order of performance of composite domains by percentages in PL 5 and 6, in descending order, was Comprehension, Oral, Overall, and Literacy.

1.3.2.1 Oral Composite

1.3.2.1.1 By Cluster

Table 1.3.2.1.1.1

Proficiency Level by Cluster (Count): Oral, S601 Online

	Oral Language Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	34866	35930	63538	45261	17389	1721	198705
2–3	57470	76770	126665	92529	30959	3607	388000
4–5	21891	25913	59493	110604	91556	31538	340995
6–8	44093	61052	123994	118639	29031	4084	380893
9–12	103114	83709	149777	70098	7891	1062	415651

Table 1.3.2.1.1.2

Proficiency Level by Cluster (Percent): Oral, S601 Online

	Oral Language Proficiency Range						
Cluster	1	2	3	4	5	6	Total
1	17.50%	18.10%	32.00%	22.80%	8.80%	0.90%	100.00%
2–3	14.80%	19.80%	32.60%	23.80%	8.00%	0.90%	100.00%
4–5	6.40%	7.60%	17.40%	32.40%	26.80%	9.20%	100.00%
6–8	11.60%	16.00%	32.60%	31.10%	7.60%	1.10%	100.00%
9–12	24.80%	20.10%	36.00%	16.90%	1.90%	0.30%	100.00%

1.3.2.1.2 By Grade

Table 1.3.2.1.2.1

Proficiency Level by Grade (Count): Oral, S601 Online

	Oral Language Proficiency Range						
Grade	1	2	3	4	5	6	Total
1	34866	35930	63538	45261	17389	1721	198705
2	28686	43442	62748	43086	13912	1615	193489
3	28784	33328	63917	49443	17047	1992	194511
4	9541	13765	32023	58687	51954	20075	186045

Grade	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
5	12350	12148	27470	51917	39602	11463	154950
6	10789	19791	43228	41939	11674	1278	128699
7	14877	21205	41070	40516	9660	1431	128759
8	18427	20056	39696	36184	7697	1375	123435
9	35731	29881	49714	25635	2763	339	144063
10	29559	23367	41362	20361	2281	328	117258
11	19965	16566	31183	14253	1811	250	84028
12	17859	13895	27518	9849	1036	145	70302

Table 1.3.2.1.2.2

Proficiency Level by Grade (Percent): Oral, S601 Online

Grade	**Oral Language Proficiency Range**						**Total**
	1	**2**	**3**	**4**	**5**	**6**	
1	17.50%	18.10%	32.00%	22.80%	8.80%	0.90%	100.00%
2	14.80%	22.50%	32.40%	22.30%	7.20%	0.80%	100.00%
3	14.80%	17.10%	32.90%	25.40%	8.80%	1.00%	100.00%
4	5.10%	7.40%	17.20%	31.50%	27.90%	10.80%	100.00%
5	8.00%	7.80%	17.70%	33.50%	25.60%	7.40%	100.00%
6	8.40%	15.40%	33.60%	32.60%	9.10%	1.00%	100.00%
7	11.60%	16.50%	31.90%	31.50%	7.50%	1.10%	100.00%
8	14.90%	16.20%	32.20%	29.30%	6.20%	1.10%	100.00%
9	24.80%	20.70%	34.50%	17.80%	1.90%	0.20%	100.00%
10	25.20%	19.90%	35.30%	17.40%	1.90%	0.30%	100.00%
11	23.80%	19.70%	37.10%	17.00%	2.20%	0.30%	100.00%
12	25.40%	19.80%	39.10%	14.00%	1.50%	0.20%	100.00%

1.3.2.2 Literacy Composite

1.3.2.2.1 By Cluster

Table 1.3.2.2.1.1

Proficiency Level by Cluster (Count): Literacy, S601 Online

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	79072	84332	45354	8427	1524	120	218829

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
2–3	57825	112545	182477	63204	6621	495	423167
4–5	43499	43006	145302	96700	17980	3361	349848
6–8	77889	102695	180364	37605	1861	71	400485
9–12	62830	109184	177394	65241	9899	436	424984

Table 1.3.2.2.1.2

Proficiency Level by Cluster (Percent): Literacy, S601 Online

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	36.10%	38.50%	20.70%	3.90%	0.70%	0.10%	100.00%
2–3	13.70%	26.60%	43.10%	14.90%	1.60%	0.10%	100.00%
4–5	12.40%	12.30%	41.50%	27.60%	5.10%	1.00%	100.00%
6–8	19.40%	25.60%	45.00%	9.40%	0.50%	0.00%	100.00%
9–12	14.80%	25.70%	41.70%	15.40%	2.30%	0.10%	100.00%

1.3.2.2.2 By Grade

Proficiency Level by **Table 1.3.2.2.2.1**

Proficiency Level by Grade (Count): Literacy, S601 Online

Grade	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	79072	84332	45354	8427	1524	120	218829
2	28659	65067	92978	24242	2100	182	213228
3	29166	47478	89499	38962	4521	313	209939
4	22540	23284	81071	51986	9761	2092	190734
5	20959	19722	64231	44714	8219	1269	159114
6	24615	35880	64614	10206	427	37	135779
7	25625	35093	61511	12972	726	18	135945
8	27649	31722	54239	14427	708	16	128761
9	21406	34861	63047	24648	3605	236	147803
10	16815	30759	49804	19295	2887	127	119687
11	12039	22805	35814	13357	2238	63	86316
12	12570	20759	28729	7941	1169	10	71178

Table 1.3.2.2.2.2

Proficiency Level by Grade (Percent): Literacy, S601 Online

Grade	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	36.10%	38.50%	20.70%	3.90%	0.70%	0.10%	100.00%
2	13.40%	30.50%	43.60%	11.40%	1.00%	0.10%	100.00%
3	13.90%	22.60%	42.60%	18.60%	2.20%	0.10%	100.00%
4	11.80%	12.20%	42.50%	27.30%	5.10%	1.10%	100.00%
5	13.20%	12.40%	40.40%	28.10%	5.20%	0.80%	100.00%
6	18.10%	26.40%	47.60%	7.50%	0.30%	0.00%	100.00%
7	18.80%	25.80%	45.20%	9.50%	0.50%	0.00%	100.00%
8	21.50%	24.60%	42.10%	11.20%	0.50%	0.00%	100.00%
9	14.50%	23.60%	42.70%	16.70%	2.40%	0.20%	100.00%
10	14.00%	25.70%	41.60%	16.10%	2.40%	0.10%	100.00%
11	13.90%	26.40%	41.50%	15.50%	2.60%	0.10%	100.00%
12	17.70%	29.20%	40.40%	11.20%	1.60%	0.00%	100.00%

1.3.2.3 Comprehension Composite

1.3.2.3.1 By Cluster

Table 1.3.2.3.1.1

Proficiency Level by Cluster (Count): Comprehension, S601 Online

Cluster	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	25296	49808	59706	24854	28370	18265	206299
2–3	41524	94173	100209	49626	61163	51112	397807
4–5	20972	46177	57458	46560	82441	90223	343831
6–8	56279	95857	92720	57800	56658	30276	389590
9–12	63534	104081	93110	57654	58047	40470	416896

Table 1.3.2.3.1.2

Proficiency Level by Cluster (Percent): Comprehension, S601 Online

Cluster	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	12.30%	24.10%	28.90%	12.00%	13.80%	8.90%	100.00%
2–3	10.40%	23.70%	25.20%	12.50%	15.40%	12.80%	100.00%
4–5	6.10%	13.40%	16.70%	13.50%	24.00%	26.20%	100.00%
6–8	14.40%	24.60%	23.80%	14.80%	14.50%	7.80%	100.00%
9–12	15.20%	25.00%	22.30%	13.80%	13.90%	9.70%	100.00%

1.3.2.3.2 By Grade

Table 1.3.2.3.2.1

Proficiency Level by Grade (Count): Comprehension, S601 Online

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	25296	49808	59706	24854	28370	18265	206299
2	13878	48522	56540	27922	30778	21688	199328
3	27646	45651	43669	21704	30385	29424	198479
4	7487	24849	31749	24767	46278	52413	187543
5	13485	21328	25709	21793	36163	37810	156288
6	14219	34079	35251	20195	19358	8503	131605
7	19096	32277	31003	20196	18653	10796	132021
8	22964	29501	26466	17409	18647	10977	125964
9	19796	37464	32142	20813	21182	13012	144409
10	18090	29018	26126	16312	15925	11984	117455
11	13248	20289	18847	11091	12337	9178	84990
12	12400	17310	15995	9438	8603	6296	70042

Table 1.3.2.3.2.2

Proficiency Level by Grade (Percent): Comprehension, S601 Online

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	12.30%	24.10%	28.90%	12.00%	13.80%	8.90%	100.00%
2	7.00%	24.30%	28.40%	14.00%	15.40%	10.90%	100.00%
3	13.90%	23.00%	22.00%	10.90%	15.30%	14.80%	100.00%

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
4	4.00%	13.20%	16.90%	13.20%	24.70%	27.90%	100.00%
5	8.60%	13.60%	16.40%	13.90%	23.10%	24.20%	100.00%
6	10.80%	25.90%	26.80%	15.30%	14.70%	6.50%	100.00%
7	14.50%	24.40%	23.50%	15.30%	14.10%	8.20%	100.00%
8	18.20%	23.40%	21.00%	13.80%	14.80%	8.70%	100.00%
9	13.70%	25.90%	22.30%	14.40%	14.70%	9.00%	100.00%
10	15.40%	24.70%	22.20%	13.90%	13.60%	10.20%	100.00%
11	15.60%	23.90%	22.20%	13.00%	14.50%	10.80%	100.00%
12	17.70%	24.70%	22.80%	13.50%	12.30%	9.00%	100.00%

1.3.2.4 Overall Composite

1.3.2.4.1 By Cluster

Table 1.3.2.4.1.1

Proficiency Level by Cluster (Count): Overall, S601 Online

Cluster	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	44722	72324	62516	11386	1814	89	192851
2–3	49467	92030	157646	65919	7948	322	373332
4–5	27886	34009	101146	116450	30404	3784	313679
6–8	52539	76682	160876	57749	3016	101	350963
9–12	70089	89477	159387	56021	5831	227	381032

Table 1.3.2.4.1.2

Proficiency Level by Cluster (Percent): Overall, S601 Online

Cluster	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	23.20%	37.50%	32.40%	5.90%	0.90%	0.00%	100.00%
2–3	13.30%	24.70%	42.20%	17.70%	2.10%	0.10%	100.00%
4–5	8.90%	10.80%	32.20%	37.10%	9.70%	1.20%	100.00%
6–8	15.00%	21.80%	45.80%	16.50%	0.90%	0.00%	100.00%
9–12	18.40%	23.50%	41.80%	14.70%	1.50%	0.10%	100.00%

1.3.2.4.2 By Grade

Table 1.3.2.4.2.1

Proficiency Level by Grade (Count): Overall, S601 Online

Grade	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	44722	72324	62516	11386	1814	89	192851
2	24373	52304	80456	26229	2721	151	186234
3	25094	39726	77190	39690	5227	171	187098
4	13386	18042	55888	63232	17590	2558	170696
5	14500	15967	45258	53218	12814	1226	142983
6	14423	26499	58653	17797	745	43	118160
7	17815	25719	54052	20187	1096	23	118892
8	20301	24464	48171	19765	1175	35	113911
9	23730	28830	55941	21004	2149	111	131765
10	19601	24966	44736	16248	1677	56	107284
11	13754	18111	32516	11372	1298	51	77102
12	13004	17570	26194	7397	707	9	64881

Table 1.3.2.4.2.2

Proficiency Level by Grade (Percent): Overall, S601 Online

Grade	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	23.20%	37.50%	32.40%	5.90%	0.90%	0.00%	100.00%
2	13.10%	28.10%	43.20%	14.10%	1.50%	0.10%	100.00%
3	13.40%	21.20%	41.30%	21.20%	2.80%	0.10%	100.00%
4	7.80%	10.60%	32.70%	37.00%	10.30%	1.50%	100.00%
5	10.10%	11.20%	31.70%	37.20%	9.00%	0.90%	100.00%
6	12.20%	22.40%	49.60%	15.10%	0.60%	0.00%	100.00%
7	15.00%	21.60%	45.50%	17.00%	0.90%	0.00%	100.00%
8	17.80%	21.50%	42.30%	17.40%	1.00%	0.00%	100.00%
9	18.00%	21.90%	42.50%	15.90%	1.60%	0.10%	100.00%
10	18.30%	23.30%	41.70%	15.10%	1.60%	0.10%	100.00%
11	17.80%	23.50%	42.20%	14.70%	1.70%	0.10%	100.00%
12	20.00%	27.10%	40.40%	11.40%	1.10%	0.00%	100.00%

2. Analysis of Domains

The measurement model that forms the basis of the analysis for the development of ACCESS for ELLs is the Rasch measurement model (Wright & Stone, 1979). Additional information on its use in the development of the ACCESS for ELLs assessment program is available in WIDA Consortium Technical Report No. 1, *Development and Field Test of ACCESS for ELLs* (Kenyon, 2006). The original ACCESS test developers used Rasch measurement principles, and in that sense, the Rasch model guided all decisions throughout the development of the assessment and was not just a tool for the statistical analysis of the data. Thus, for example, data based on Rasch fit statistics guided the inclusion, revision, or deletion of items during the development and field testing of the test forms and will continue to guide the refinement and further development of the test. All Rasch analyses are conducted using the Rasch measurement software program *Winsteps* (Linacre, 2006).

Rasch Model for Dichotomous Scoring

For Listening and Reading, the dichotomous Rasch model was used as the measurement model. Mathematically, the measurement model may be presented as

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

where

P_{ni1} = probability of providing a correct response "1" by student "n" to item "i"

P_{ni0} = probability of providing an incorrect response "0" by student "n" to item "i"

B_n = ability of student "n"

D_i = difficulty of item "i"

When the probability of a student providing a correct answer to an item equals the probability of a student providing an incorrect answer (i.e., 50% probability of getting it right and 50% probability of getting it wrong), P_{ni1}/P_{ni0} is equal to 1. The log of 1 is 0. This is the point at which a student's ability equals the difficulty of an item. For example, a student whose ability estimate is 1.56 on the Rasch logit scale encountering an item whose difficulty is 1.56 on the Rasch logit scale would have a 50% probability of providing a correct answer to that item.

Rasch Model for Polytomous Scoring

The Writing and Speaking tasks used a Rasch-grouped rating scale model, which is an extension of Andrich's rating scale model (Andrich, 1978). Mathematically, this can be represented as

$$\log\left(\frac{P_{ngik}}{P_{ngi(k-1)}}\right) = \beta_n - D_{gi} - F_{gk}$$

where

P_{ngik} = probability of student "n" on task "i" receiving a rating at level "k" on rating scale "g"

$P_{ngi(k-1)}$ = probability of student "n" on task "i" receiving a rating at level "k – 1" on rating scale "g" (i.e., the next lowest rating)

β_n = ability of student "n"

D_{gi} = difficulty of task "i" specific to rating scale "g"

F_{gk} = step calibration value of category "*k*" relative to category '*k* – 1' on rating scale "g"

The subscript "g" is a group index specifying the group of tasks to which task "i" belongs. It also identifies the rating scale that was used for the group of tasks. There is only one rating scale (g = 1) in the Writing domain and two grouped rating scales (g = 2) in the Speaking domain. As with the dichotomous Rasch model, there is an item difficulty parameter (D_{gi}) for each item for rating scale "g" modeled by the Rasch rating scale model (Andrich, 1978). In addition, there is a step calibration value or *step measure* (F_{gk}) that corresponds to the location on the latent variable where the probability of being observed in the "k" and "k – 1" category for rating scale "g*"* is equal, relative to the difficulty measure of the task. The step measures are also the points where adjacent category probability "k – 1" and "k" curves for rating scale "g" intercept. All tasks that belong to the same rating scale group have the same step measures. As described in Part 1 Section 3.2.3, ratings on the ACCESS Writing Scoring Scale range from 0, 1, 1+, …, 6, and the possible raw scores range from 0 to 9. Writing raters use this scoring scale for all Writing tasks. We model all other Writing tasks using a single rating scale with possible raw scores of 0 to 9.

In 2015–2016, with the transition to Online ACCESS, CAL conducted a Writing scaling study. Detailed information about the derivation of the Writing rating scale as well as the psychometric properties of the Writing rating scale are available in the 2016 scaling report (Center for Applied

Linguistics, 2017). In 2019–2020, we redesigned the Writing test to allow for embedded field testing, reducing the number of operational tasks from three to two. For details on how we retained the 2016 rating scale parameters and maintained the Writing score scale, see Center for Applied Linguistics (2019).

For Speaking, we model PL 1 tasks as a group on a 0–2 scale, and PL 3 and PL 5 tasks as a group on a 0–4 scale (see Part 1 Section 3.2.4). We conducted a study in the summer of 2016 to reconstruct the logit scales, and detailed information about the derivation as well as the psychometric properties of Speaking rating scales are available in the scaling report (Center for Applied Linguistics, 2017).

Scale Scores and Proficiency Level Scores

Scale scores are calculated by transforming the student ability estimate via a scaling equation. The following scaling equations convert ability measures in logits to scale scores:

L: (Ability Measure in Logits * 37.571) + 316.637

R: (Ability Measure in Logits * 26.000) + 323.272

W: (Ability Measure in Logits * 26.851) + 303.332

S: (Ability Measure in Logits * 29.248) + 265.076

In the domains of Listening and Reading, we established the current ACCESS scale for the original paper-only version of the test and maintained this scale through the transition to an online- and paper-delivered test in the 2015–2016 school year (Series 400). Evidence for scale maintenance in the transitional year is described elsewhere (Center for Applied Linguistics, 2016). In the domains of Writing and Speaking, we conducted a study in the summer of 2016 to reconstruct the logit scale (Center for Applied Linguistics, 2017).

PL scores are interpretations of these scale scores in terms of the proficiency levels described in the WIDA ELD Standards. These interpretations derive from a series of standard-setting studies, in which educators reviewed evidence from the test, either in the form of items for the selected response sections (Listening and Reading) or student portfolios for the constructed response sections (Writing and Speaking), to establish cut scores between the proficiency levels. The first standard-setting study for ACCESS took place in 2005; it established cut scores for all four

domains by grade-level cluster (Kenyon, 2006). The second cut score study took place in 2007; it established cut scores for all four domains by grade level (Kenyon, Ryu, & MacGregor, 2013). These cut scores were used to derive proficiency level scores through the 2015–2016 administration (Series 400) of ACCESS for ELLs. WIDA and CAL conducted a third cut score study in summer 2016 (Cook & MacGregor, 2017). The purpose of this study was to re-examine cut scores for each of the proficiency levels in light of the migration from the paper-and-pencil–only assessment to both online and paper delivery, the revision of the Speaking test, and the influence of college- and career-ready standards. These new cut scores were first used for ACCESS Series 401 (2016–2017 school year).

A proficiency level score consists of a two-digit decimal number (e.g., 4.5). The first digit represents the student's overall proficiency level range based on the student's scale score. The number to the right of the decimal is an indication of the proportion of the range between cut scores that the student's scale score represents. A score of 4.5, for example, tells us that the student is in PL 4 and that the student's scale score is halfway between the cut scores for PLs 4 and 5.

Unlike the scale scores, which form an interval scale and are continuous across grades from Kindergarten to Grade 12, PL scores are dependent upon the grade a student was in when the student took the assessment. For example, a score of 350 in Listening would be interpreted as a PL score of 5.8 for a Grade 2 student, a 3.8 for a Grade 5 student, a 3.1 for a Grade 8 student, and a 2.3 for a Grade 12 student.

Because the bands between cut scores on the score scale vary in width, PL scores do not form an interval scale. Only scale scores should be used as interval measures. PL scores are at even intervals within a grade and proficiency level (e.g., in Grade 3, the distance between 3.1 and 3.2 is the same as the distance between 3.7 and 3.8), but they do not form an interval scale across proficiency levels.

2.1 Complete Item or Task Analysis and Summary

The tables in this section provide information on the psychometric qualities of the items and tasks. We provide values for item or task difficulties in logits, the number of items or tasks on the form, the average *p* value (for forms with selected response items), and the Rasch model fit statistics. For Writing and Speaking, we also provide raw score distributions by task.

Tables in this section have either two parts (in the case of Listening and Reading) or three parts (in the case of Writing and Speaking). The first part of the table gives a summary of the total set of items or tasks on the form. The second part provides statistics pertaining to the individual items or tasks, and the third part (for Writing and Speaking only) expresses raw score distributions by task.

For Listening and Reading, items form a pool for the multistage adaptive tests, and tables in this section provide information on every item in the grade-level cluster. For Writing, separate tables are provided for Tier A and Tier B/C forms, by grade-level cluster. For Speaking, which has tasks that are shared between Tier A and Tier B/C, there is one table for each grade-level cluster, which provides information on every task in the grade-level cluster.

All Rasch analyses were conducted using the Rasch measurement software program *Winsteps 4.8.2.0* (Linacre, 2006). When speaking of the measure of student ability, we use the term *ability measure* (rather than *theta*, used commonly when discussing models based on item response theory). When speaking of the measure of how hard an item is, we use the term *item difficulty measure* (rather than *b parameter*, used commonly when discussing models based on item response theory). *Step measures* refer to the calibration of the steps in the Rasch rating scale model previously presented. All three measures (ability, difficulty, and step) are expressed in terms of Rasch logits, which then are converted into scores on the ACCESS score scale for reporting purposes.

Fit statistics for the Rasch model are calculated by comparing the observed empirical data with the data that the Rasch model would be expected to produce if the data fit the model perfectly. Outfit mean square statistics for items and tasks are influenced by outlier responses for machine-scored dichotomous items or outlier ratings for rater-scored performance tasks. For example, a difficult item that some low-ability students get correct—for reasons unknown—will have a high

outfit mean square statistic. Similarly, an easy item that some high-ability students get wrong will also have a high outfit mean square statistic. Infit mean square statistics are influenced by unexpected patterns of students' responses and ratings on items and tasks that are roughly targeted for them and generally indicate a more serious measurement problem. The expectation for both statistics is 1.00, and values near 1.00 are not of great concern. Values less than 1.00 indicate that the response and rating patterns are too predictable and thus redundant, or the model is overfitting the data, but are not of great concern. High values are of greater concern.

Linacre (2002b) provided more guidance on how to interpret these statistics for dichotomous items. He wrote:

- Values greater than 2.0 "distort or degrade[1] the measurement system."
- Values between 1.5 and 2.0 are "unproductive for construction of measurement, but not degrading."
- Values between 0.5 and 1.5 should be considered "productive for measurement."
- Values below 0.5 are "less productive for measurement, but not degrading."

Linacre also stated in his guidance that infit problems are more serious to the construction of measurement than are outfit problems.

Because we follow conservative guidelines in the development of ACCESS for ELLs, it is desired that the dichotomous items on the test forms have mean square fit statistics in the range of 0.5 to 1.5; and thus, they fit the range that is "productive for measurement" according to the guidelines above. See below for the percentages of dichotomous items which have mean square statistics within this range, by domain.

Since performance tasks are constructed and scored very differently from dichotomous items, it is not as straightforward to apply this same guidance to interpret these fit statistics for performance tasks that raters scored polytomously on a rubric scale. We design some performance tasks to elicit a restricted range of performances (for example, very easy tasks where we expect that most students will get the highest rating), and these tasks can cause the model to predict the data too well (overfitting). Conversely, when raters score performance tasks using a very wide rubric scale such as the ACCESS for ELLs Writing rubric, sometimes

[1] We interpret "degrade" here in the sense of lowering the quality of the measurement system.

unmodeled noise or other sources of variance in the ratings of the students' responses to the task will cause the model to underpredict those ratings (underfitting). Overall, for ACCESS for ELLs performance tasks, overfitting is more common than underfitting. Underfitting indicates that the task is less productive for measurement, but, according to Linacre (2002b), including the rating of the student's performance on the task when calculating that student's score does not degrade the measurement of the student's performance.

The first section of the Complete Item/Task Analysis and Summary table provides information about the total set of items or tasks and includes the item type (selected response or constructed response), the average item difficulty measure (in logits), the number of items, the average *p* value (for Listening and Reading only), the average infit mean square statistic, and the average outfit mean square statistic.

The second section of these tables presents results from the analyses of all the items or tasks on the test form. The first column provides the unique item name. The second column in this section presents the item or task difficulty measure in logits. The third column indicates whether the item (or task) served as an anchor item (or task) which are used to link score scales between Series (See 2.7, Equating Summary for details) or dichotomously scored items (Listening and Reading), the fourth column shows the *p* value (percentage of correct answers on that item). The final two columns show the Rasch fit statistics for the item or task. Folders with items that have fit statistics greater than 2.0 are evaluated by the test development team to determine whether and when the folders can be refreshed in the next test refreshment cycle.

In addition, Writing and Speaking tables have a section at the bottom of the table that provides raw score distributions by task.

The results show that nearly all items and tasks have infit mean square statistics less than 2.0 for all grade-level clusters and domains, indicating that the items and tasks provide trustworthy measures of ability for those students whose ability measures are in the region of the ability distribution that the items and tasks are targeting. As discussed earlier, the outfit mean square statistic is sensitive to outlier responses and scores that are not in the region of the ability distribution that the items and tasks are targeting. There are two items in the Grades 2–3 Listening test that show outfit mean square statistics greater than 2.0. For the most part, these are

very easy items, suggesting that there might be some high-ability students getting these items incorrect and causing the outfit mean square statistics to be inflated.

All items in the Listening and Reading domains have infit mean square statistics between 0.5 and 1.5. All items in the Listening clusters 01 and 45, and all Reading clusters except cluster 912 have outfit mean square statistics that fall between 0.5 and 1.5. Listening clusters 23, 68, 912, and Reading cluster 912 have slightly lower outfit mean square statistics, with 94%, 98%, 98%, and 99% falling between 0.5 and 1.5, respectively.

Note: The redacted version of the annual technical report does not have item related information (tables are removed from section 2.1 Complete Item or Task Analysis and Summary and section 2.2 DIF Analysis and Summary).

2.2 DIF Analysis and Summary

Prior to field testing, the Bias and Sensitivity Review Panel ensures that test items and tasks are free of material that (1) might favor any subgroup of students over another on the basis on gender, race/ethnicity, home language, religion, culture, region, or socioeconomic status, and (2) might be upsetting to students. This process is qualitatively driven, while the DIF analysis, described below, is data driven. Please see Part 1, section 2.3.1 for more information on Bias and Sensitivity panels.

CAL uses differential item functioning (DIF) analysis to investigate whether factors extraneous to English language proficiency (i.e., the construct being measured on the test) may have influenced some students' performances on items. DIF attempts to find items that may be functioning differently for different groups based on criteria irrelevant to the construct that is purportedly being measured. We compare the performance of students on ACCESS for ELLs Online items and tasks by dividing students into two different groupings: first, males versus females; second, students of Hispanic ethnic background versus students of all other backgrounds. For the former analysis, females is the reference group, while males are the focal group. For the latter analysis, Hispanics is the reference group, while Non-Hispanics is the focal group. We exclude students for whom gender or ethnicity[2] was unknown from both analyses. We used two commonly used procedures for detecting DIF: one for dichotomously scored items (Listening and Reading), conducted prior to operational testing, and one for polytomously scored items (Writing and Speaking), conducted on population data after the close of operational testing.

Dichotomous Items

We used the Mantel-Haenszel (M-H) chi-square statistic (Mantel & Haenszel, 1959) procedure for dichotomous items, originally proposed by the Educational Testing Service (ETS). This procedure compares item-level performances of students in the two groups (e.g., males versus females) who are divided into subgroups based on their performance on the total test. We assume that if there is no DIF, a similar percentage of students in each group should get the item correct at any ability level (based on performance on the total test). We use the M-H chi-square statistic to check the probability that the two groups performed comparably on each item across the ability groupings. The statistic is transformed into the "M-H delta" scale. This scale is symmetrical around zero, with a delta zero interpreted as indicating that neither group is favored. A positive result indicates that the focal group is favored; a negative result indicates that the reference group is favored.

The existing M-H procedure was designed for fixed forms, where all students take the same set of items; therefore, the students can be matched on the number-correct score when computing

[2] In the dataset, Hispanic ethnicity, as well as each of the race categories, is coded as a binary variable (Y/blank). Ethnicity information is counted as "Unknown" in cases where the student is recorded as blank for Hispanic ethnicity and also blank for every race category.

the M-H statistic. In the multistage computerized adaptive test condition, however, not all students take the same set of items; thus, it is not possible to match students on the number-correct score. Instead, we use a computerized adaptive test M-H DIF procedure (Zwick, Thayer, & Wingersky, 1993) to examine DIF for the Listening and Reading domains. First, we derive the student's expected true score for the entire item pool. To derive the expected true score, we transform each student's Rasch ability estimate into the expected true score metric by calculating the sum of the item response functions in the operational item pool, which is evaluated at the estimated ability level of the student. We use the expected true score of the students as the matching variable for the M-H DIF procedure. Once we have matched students on the expected true score, the ordinary M-H DIF procedure and the ETS evaluation criterion for severity of M-H DIF can be applied. In CAL's implementation of this method, students are matched for M-H DIF analysis based on this expected true score using two-unit intervals, as Zwick and Bridgeman (2014) recommended. We used a two-step purification process in conducting the DIF analysis; that is, we removed items with C-level DIF in the first pass from the matching variable in the second stage, and then we recalculated the DIF for the remaining items.

Because DIF is measured on a continuous scale, and because most items are likely to show some degree of DIF, it is useful to have guidelines to determine when the level of DIF requires further review of the item. We follow the guidance provided by ETS (Zieky, 1993) to classify items into DIF levels as follows:

- A (no DIF) when the absolute value of delta is <1.0
- B (weak DIF) when the absolute value of delta is 1.0 to 1.5
- C (strong DIF) when the absolute value of the delta is >1.5

Polytomous Items

For polytomous items (i.e., Writing and Speaking tasks), we take a similar approach. Our approach is based on the M-H chi-square statistic and the standardized mean difference following procedures that ETS developed (Allen, Carlson, & Zalanak, 1999; Zwick, Donoghue, & Grima, 1993). These DIF procedures for polytomous items were used to identify tasks that exhibit DIF. We used JMetrik (Meyer, 2018), an open-source computer program for psychometric analysis, to conduct the analyses. The procedures implemented in JMetrik first

calculate the Cochran-Mantel-Haenszel chi-square statistic for testing statistical significance. This statistic gives an indication of the probability that observed differences are the result of chance but does not indicate how significant that difference is. To indicate how significant the difference is, we calculate the standardized mean difference between the performances of the two comparison groups. The standardized mean difference compares the means of the two groups, adjusting for differences in the distribution of the groups across the values of the total raw scores. To standardize the outcome, this difference is divided by the item score range and serves as an effect size measure for the Cochran-Mantel-Haenszel chi-square statistic. This effect size measure (reported as standardized P-DIF in JMetrik) ranges from -1 to 1, which may present some interpretation challenges. To mitigate the negative value, the absolute value of the Cochran-Mantel-Haenszel chi-square statistic is used in JMetrik (Meyer, 2018) and the range of the rescaled effect size (standardized P-DIF*) is restricted to fall between 0 and 1. The effect size flagging criterion for polytomous items that ETS proposed (Allen et al., 1999) is also rescaled to the standardized P-DIF* metric (Meyer, 2018).

Following guidance that ETS proposed for the National Assessment of Educational Progress (Allen et al., 1999), we classify ACCESS for ELLs Writing and Speaking tasks into three DIF levels as follows:

- AA (no DIF), when the Cochran-Mantel-Haenszel chi-square statistic is not significant or when it is significant and standardized P-DIF* is <0.05
- BB (weak DIF), when the Cochran-Mantel-Haenszel chi-square statistic is significant and standardized P-DIF* is ≥ 0.05 but <0.10
- CC (strong DIF) when the Cochran-Mantel-Haenszel chi-square statistic is significant and standardized P-DIF* is ≥ 0.10

The tables in this section provide a summary of the findings of the DIF analyses at the top, followed by information for any item or task which showed B, BB, C, or CC-level DIF. The first column gives the DIF level: A, B, or C for dichotomous items or AA, BB, or CC for polytomous tasks (i.e., Writing and Speaking tasks). The next columns show the contrasting groups in the DIF analyses: either male (focal group) versus female (reference group) or Hispanic (reference group) versus non-Hispanic other ethnicities (focal group). The top part of the table summarizes the number of items that exhibit DIF falling into each of the three categories (A, B, or C for

Listening and Reading, and AA, BB, or CC for Writing and Speaking). Any items that show B (or BB) or C (or CC)–level DIF are reported in the bottom part of the table.

If an item or task shows a C level DIF, a DIF panel is convened. The DIF panel manager, from CAL, draws panelists from CAL staff members. Members are chosen so that a diverse background is represented. Therefore, the panel manager considers gender, first/second language backgrounds, and ethnicity when empaneling judges. The manager also ensures that some members have expertise in English as a Second Language instruction and/or professional development for teachers of ESL students. Without being told which, if any, items have an initial DIF finding, the panel is asked to discuss all items in the affected folder and come to a consensus on whether they believe or do not believe that the items demonstrate bias against a particular group and is or is not appropriate to place on the operational test.

For Listening and Reading items, we conduct DIF analysis and review prior to item selection, and we remove from the item selection pool any items that the panel judges to be inappropriate. Items that exhibited a C-level DIF but were judged to have no bias by the panel can be used in future series without the need to put the item before the panel again, per WIDA's policy.

There is not sufficient scored data for DIF analysis of Speaking and Writing tasks prior to operational testing. We conduct DIF analysis using population data after operational testing is completed. Should a task exhibit CC-level DIF and should the review panel identify concern with that task, we recommend removal of the task from the subsequent year's test.

For Series 601, one item in Listening Grade 1 and one item in Listening Grades 2–3 showed C-level DIF. These items were reviewed by a panel as described above, with both Listening Grades 1 and 2–3 items being reviewed in previously held panels. These panels were not able to detect any reason for bias in the performance of these items and recommended that the items be retained on the assessment.

Note: The redacted version of the annual technical report does not have item related information (tables are removed from section 2.1 Complete Item or Task Analysis and Summary and section 2.2 DIF Analysis and Summary).

2.3 Raw Score Distribution for Speaking and Writing

Figures and tables in this section provide raw score information for Speaking and Writing. For each grade-level cluster and tier combination, the figure shows the distribution of the raw scores. The horizontal axis shows the raw scores. The vertical axis shows the number of students (count). Each bar shows how many students received each raw score.

Each table in this section summarizes results for a grade-level cluster and tier combination (e.g., Speaking 4–5 Tier A). For each table, results are broken down by grade and presented for the grade-level cluster for that tier. The following information is included in each table:

- The number of students in the analyses (the number of students who were not absent, invalid, refused, exempt, or in the wrong grade-level cluster)
- The minimum observed raw score
- The maximum observed raw score
- The mean (average) raw score
- The standard deviation (std. dev.) of the raw scores

Test design and student population impact the distribution of raw scores. In general, raw score distributions tend to be smoothly distributed with a single peak; however, there are several exceptions. Understanding these distributions supports the understanding of other statistical properties of the test forms.

Speaking Pre-A forms are designed for students at the very earliest stages of English language proficiency. Students routed to the Pre-A form have very low performances on Listening and Reading and are administered three Speaking tasks, each scored 0 to 2, for a total raw score range of 0 to 6. Tasks on the Pre-A form are by design very easy and intended to ensure beginning students are not discouraged. Large numbers of students can achieve all six points on this form. Students routed to the A form take three PL 1 tasks, scored 0 to 2, and three PL 3 tasks, scored 0 to 4, for a total raw score range of 0 to 18. Students routed to take the B/C form did not take the P1 tasks, as it is assumed that they would be able to get the full two points on these very easy PL 1 tasks. These students take three PL 3 and three PL 5 tasks, each scored 0 to 4, and they are awarded two points on each of three PL 1 tasks. The total raw score range for the Tier B/C form is 6 to 30.

2.3.1 Listening

The ACCESS 2.0 Online Listening test is a multistage adaptive assessment. As students do not all take the same set of items in the test, raw score distributions are not presented.

2.3.2 Reading

The ACCESS 2.0 Online Reading test is a multistage adaptive assessment. As students do not all take the same set of items in the test, raw score distributions are not presented.

2.3.3 Writing

2.3.3.1 Grade 1

Table 2.3.3.1.1

Raw Score Descriptive Statistics: Writ 1 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	205,653	0	14	5.21	3.06
Total	205,653	0	14	5.21	3.06

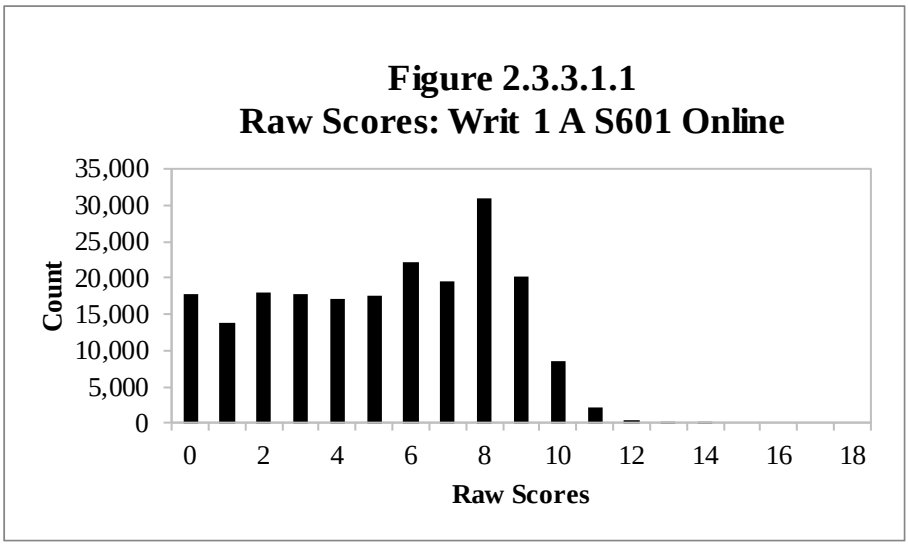
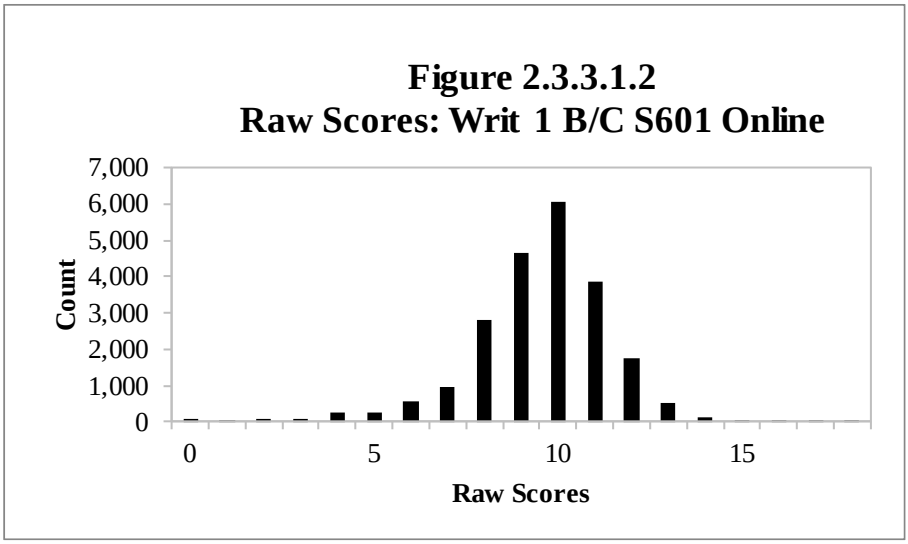


Table 2.3.3.1.2

Raw Score Descriptive Statistics: Writ 1 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	22,197	0	18	9.47	1.98
Total	22,197	0	18	9.47	1.98



2.3.3.2 Grade 2-3

Table 2.3.3.2.1

Raw Score Descriptive Statistics: Writ 2-3 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	74,987	0	15	5.17	3.22
3	62,236	0	15	5.91	3.27
Total	137,223	0	15	5.50	3.26

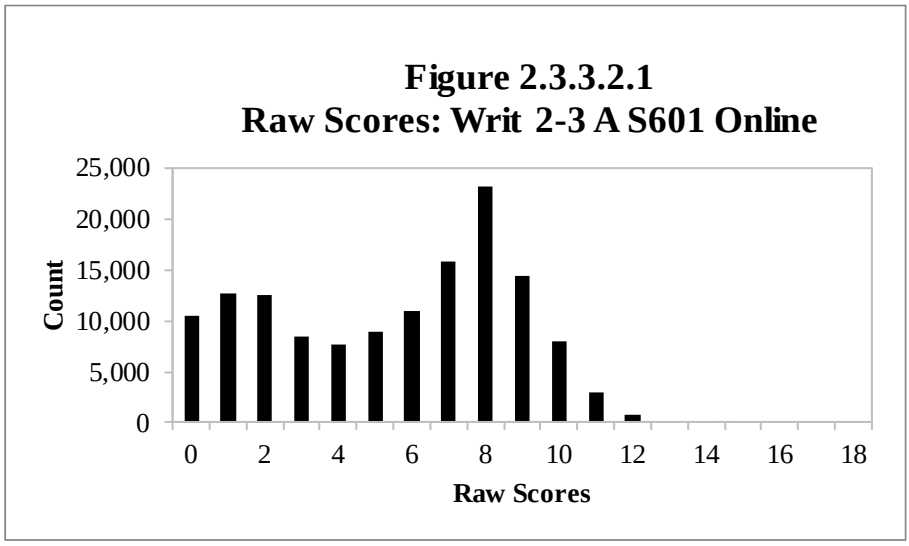
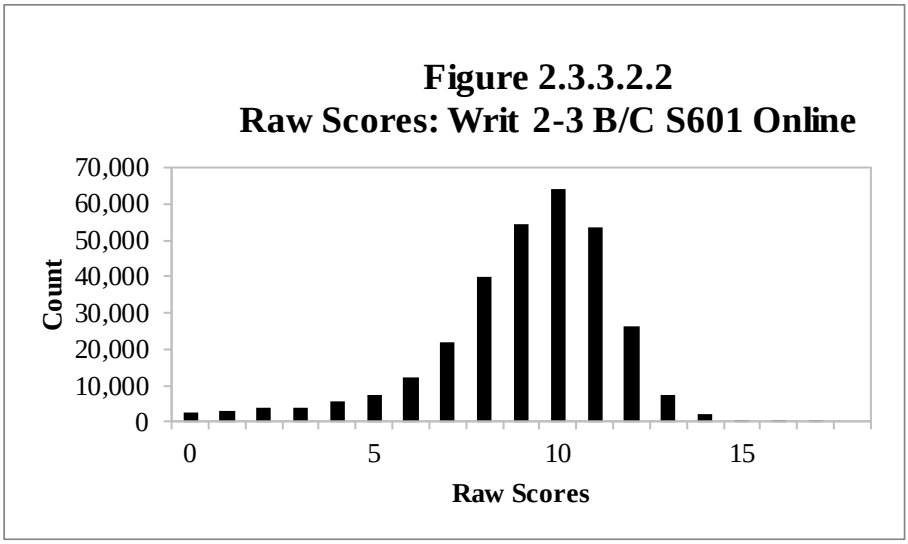


Table 2.3.3.2.2

Raw Score Descriptive Statistics: Writ 2-3 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	149,534	0	16	8.31	2.66
3	159,148	0	17	9.74	2.14
Total	308,682	0	17	9.05	2.51



2.3.3.3 Grades 4-5

Table 2.3.3.3.1

Raw Score Descriptive Statistics: Writ 4-5 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	43,240	0	15	4.29	2.99
5	43,207	0	15	5.08	3.11
Total	86,447	0	15	4.69	3.08

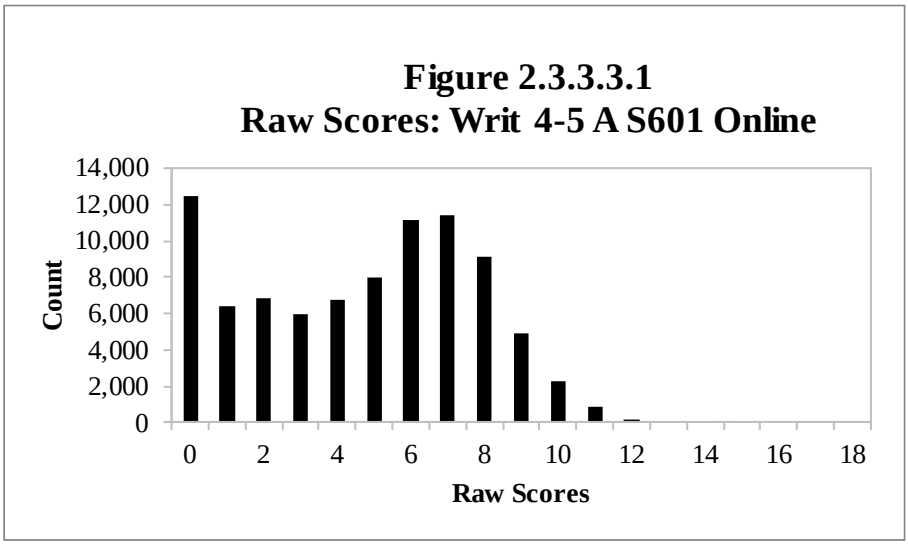
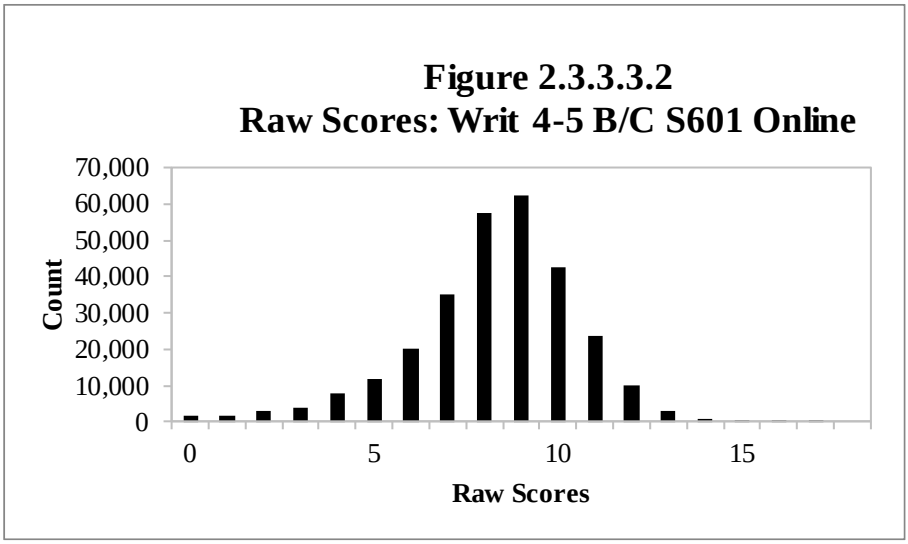


Table 2.3.3.3.2

Raw Score Descriptive Statistics: Writ 4-5 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	160,106	0	16	7.90	2.34
5	126,160	0	17	8.72	2.05
Total	286,266	0	17	8.26	2.25



2.3.3.4 Grades 6-8

Table 2.3.3.4.1
Raw Score Descriptive Statistics: Writ 6-8 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	49,195	0	16	6.24	3.05
7	60,604	0	18	6.99	3.04
8	63,844	0	18	7.47	3.05
Total	173,643	0	18	6.95	3.09

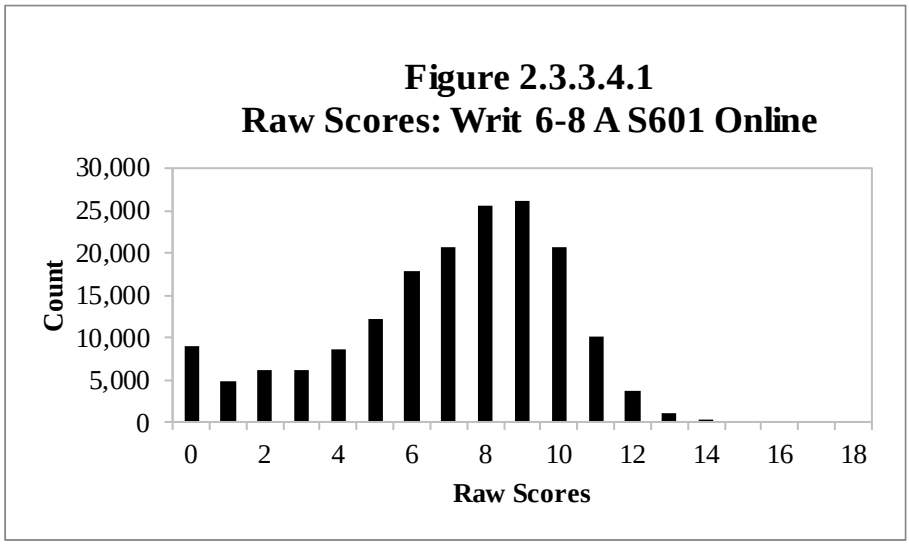
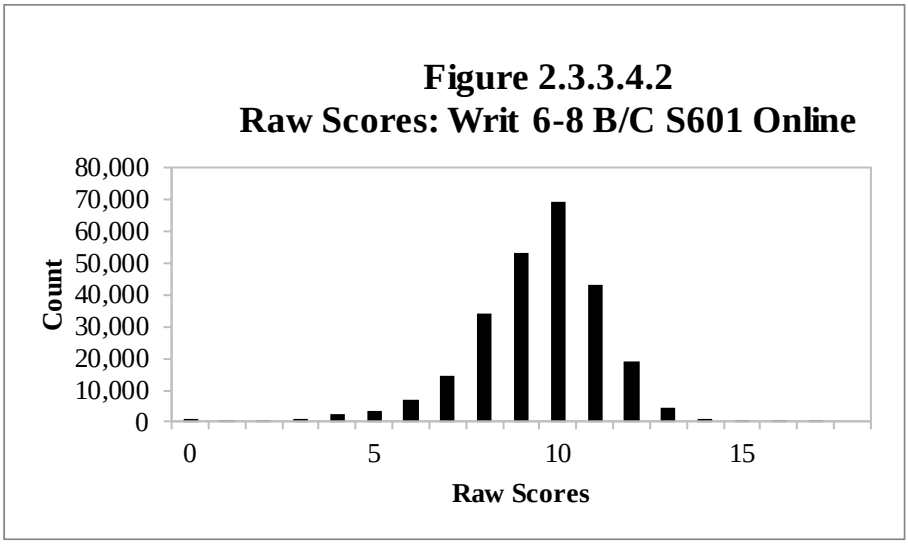


Table 2.3.3.4.2

Raw Score Descriptive Statistics: Writ 6-8 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	96,443	0	16	8.75	1.97
7	84,839	0	16	9.59	1.74
8	73,353	0	17	10.09	1.64
Total	254,635	0	17	9.42	1.89



2.3.3.5 Grades 9-12

Table 2.3.3.5.1

Raw Score Descriptive Statistics: Writ 9-12 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	66,714	0	17	5.75	3.32
10	49,466	0	18	6.26	3.01
11	32,094	0	16	6.84	2.88
12	24,237	0	16	7.01	2.88
Total	172,511	0	18	6.27	3.13

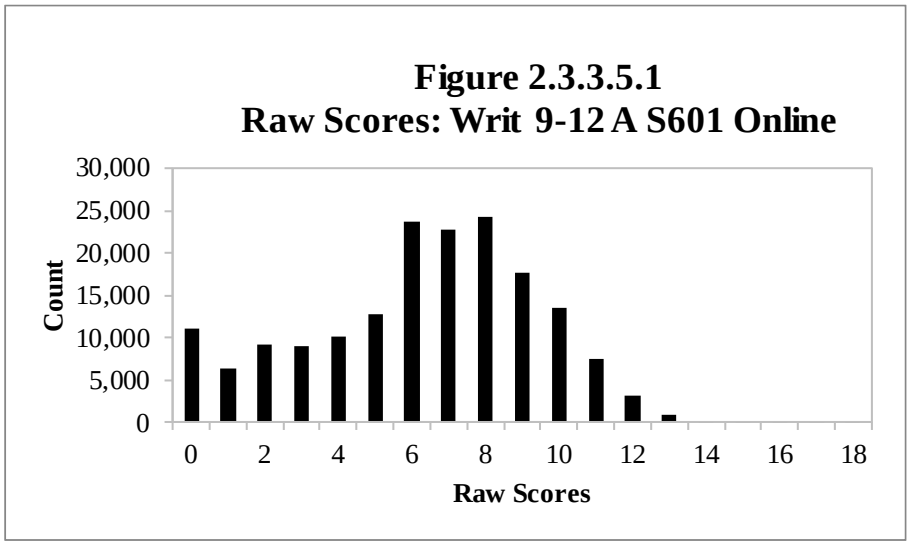
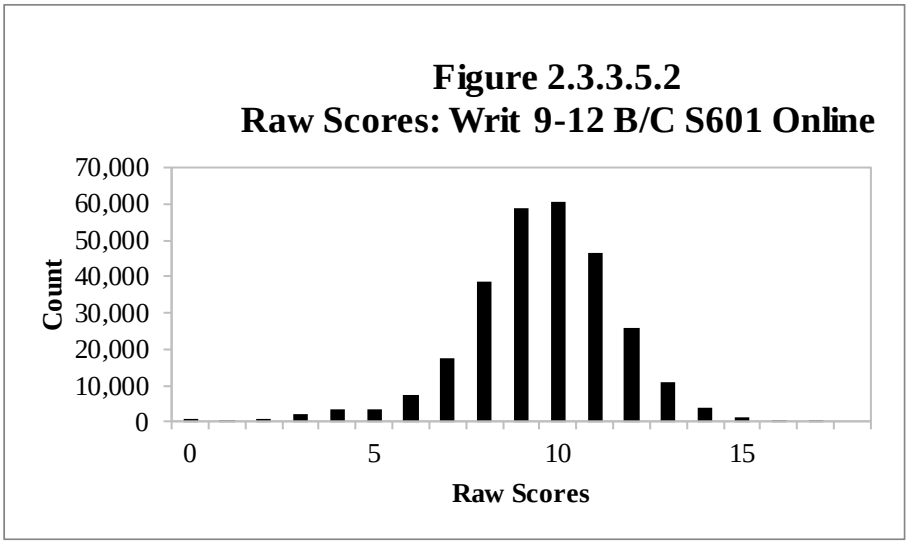


Table 2.3.3.5.2

Raw Score Descriptive Statistics: Writ 9-12 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	92,334	0	18	9.40	2.12
10	79,223	0	18	9.48	2.11
11	60,494	0	18	9.71	2.10
12	51,684	0	18	9.58	2.17
Total	283,735	0	18	9.52	2.13



2.3.4 Speaking

2.3.4.1 Grade 1

Table 2.3.4.1.1
Raw Score Descriptive Statistics: Spek 1 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	13,774	0	6	4.38	2.15
Total	13,774	0	6	4.38	2.15

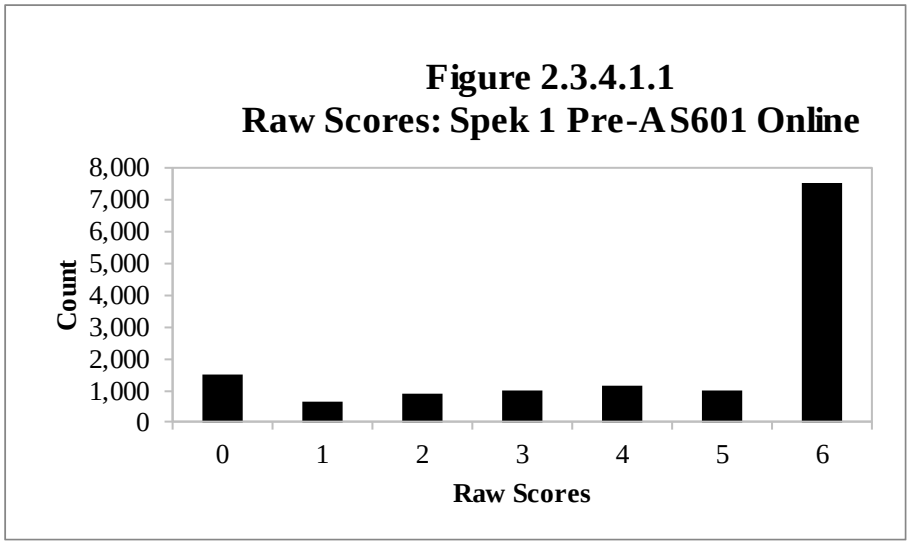


Table 2.3.4.1.2

Raw Score Descriptive Statistics: Spek 1 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	101,779	0	18	10.36	3.34
Total	101,779	0	18	10.36	3.34

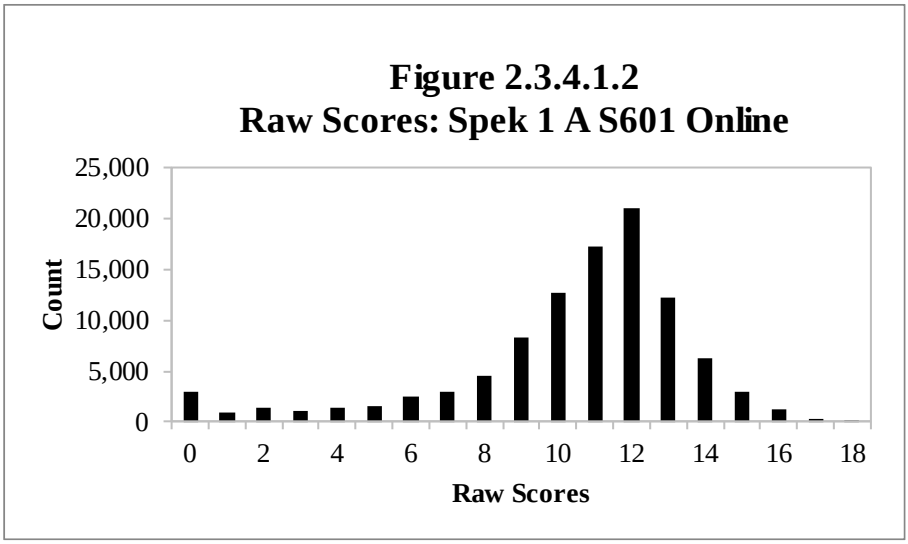
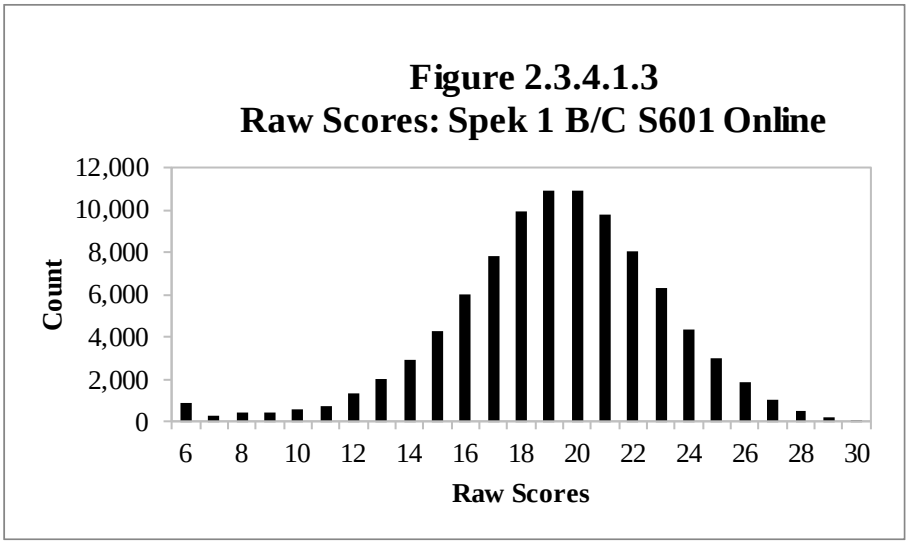


Table 2.3.4.1.3

Raw Score Descriptive Statistics: Spek 1 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	95,004	6	30	19.17	3.90
Total	95,004	6	30	19.17	3.90



2.3.4.2 Grade 2-3

Table 2.3.4.2.1

Raw Score Descriptive Statistics: Spek 2-3 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	6,222	0	6	4.62	2.06
3	14,627	0	6	4.77	1.99
Total	20,849	0	6	4.72	2.01

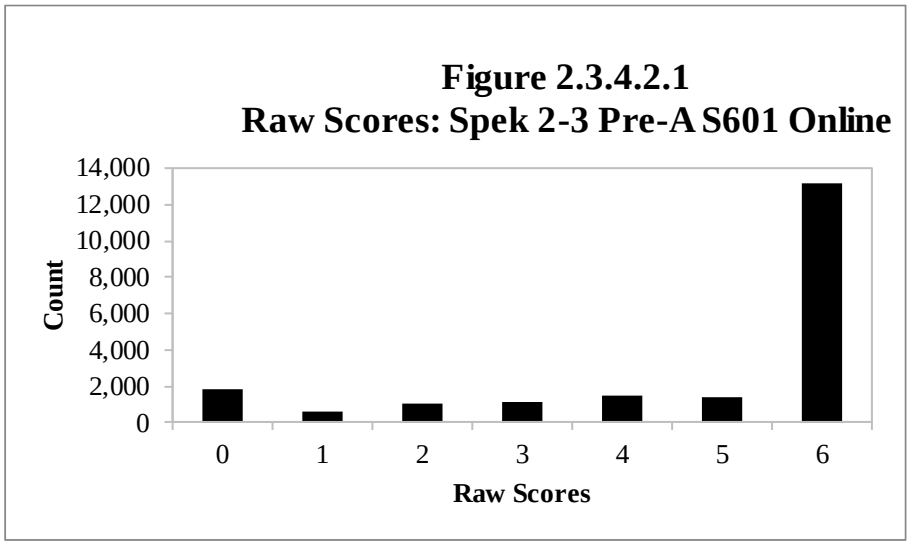


Table 2.3.4.2.2

Raw Score Descriptive Statistics: Spek 2-3 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	61,228	0	18	10.02	3.54
3	60,840	0	18	11.36	3.19
Total	122,068	0	18	10.69	3.43

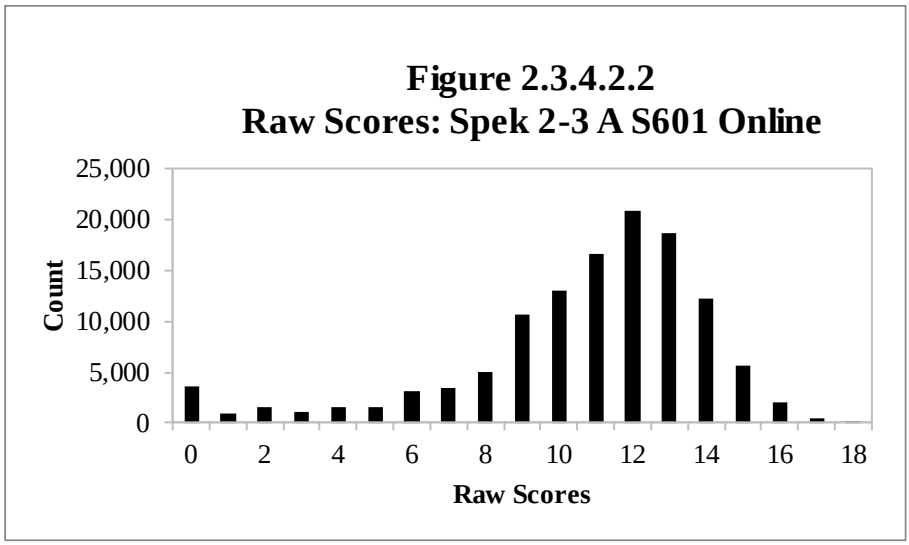
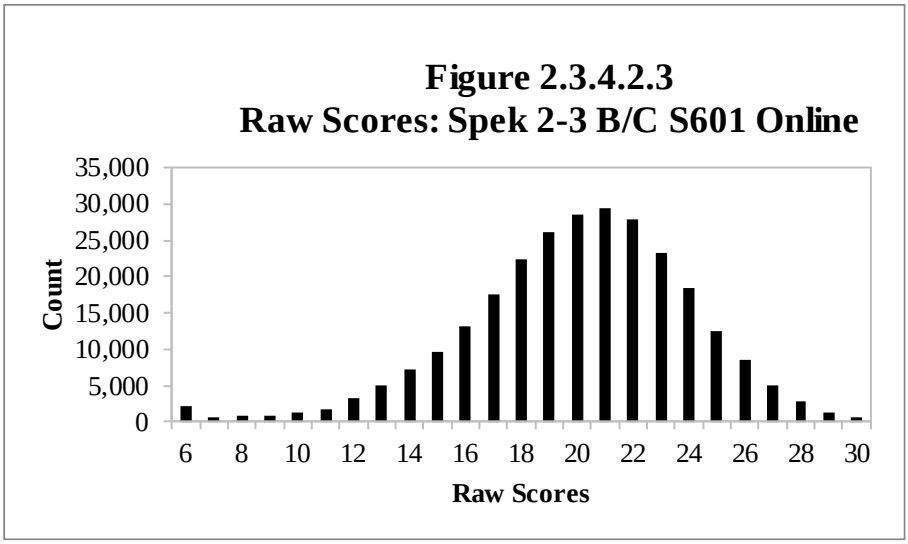


Table 2.3.4.2.3

Raw Score Descriptive Statistics: Spek 2-3 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	139,524	6	30	19.12	4.00
3	130,550	6	30	20.96	3.78
Total	270,074	6	30	20.01	4.00



2.3.4.3 Grades 4-5

Table 2.3.4.3.1

Raw Score Descriptive Statistics: Spek 4-5 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	2,426	0	6	4.26	2.13
5	6,115	0	6	4.51	2.05
Total	8,541	0	6	4.44	2.07

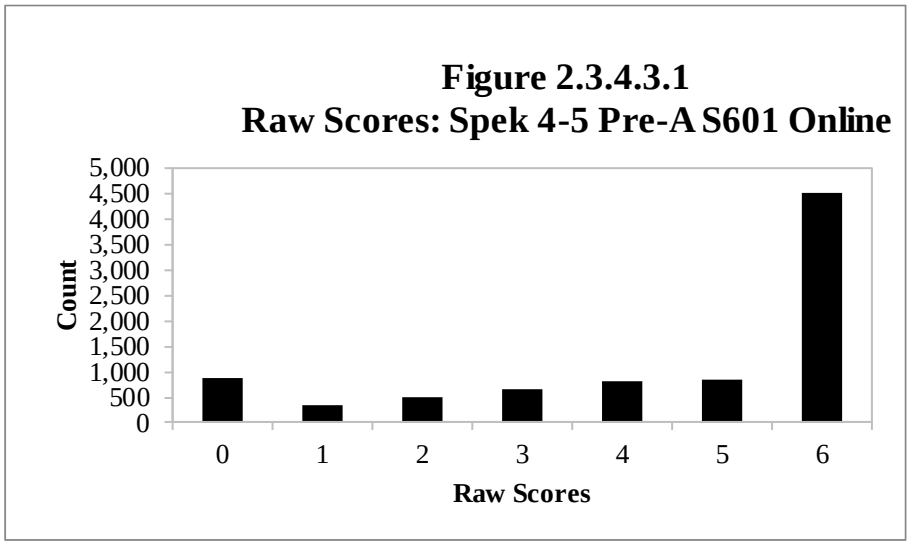


Table 2.3.4.3.2

Raw Score Descriptive Statistics: Spek 4-5 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	32,192	0	18	9.60	3.51
5	25,549	0	18	9.90	3.41
Total	57,741	0	18	9.73	3.47

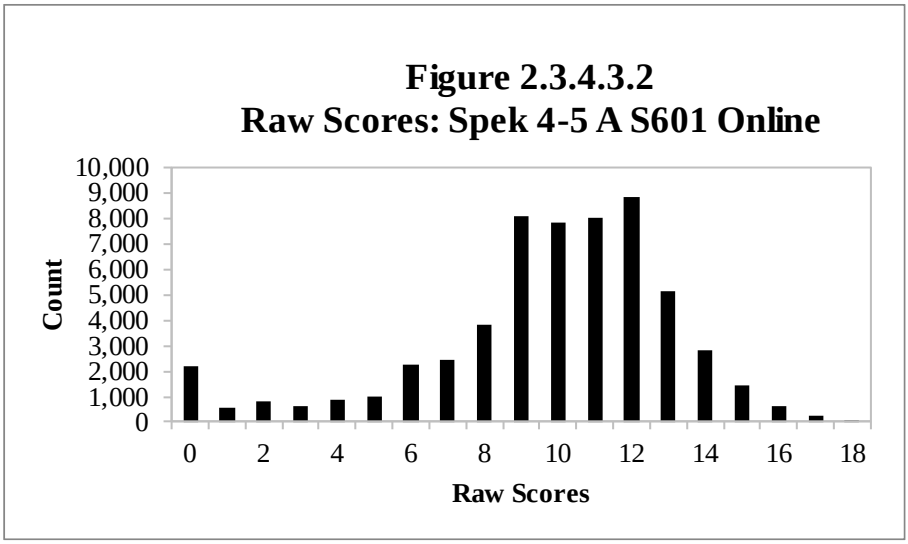
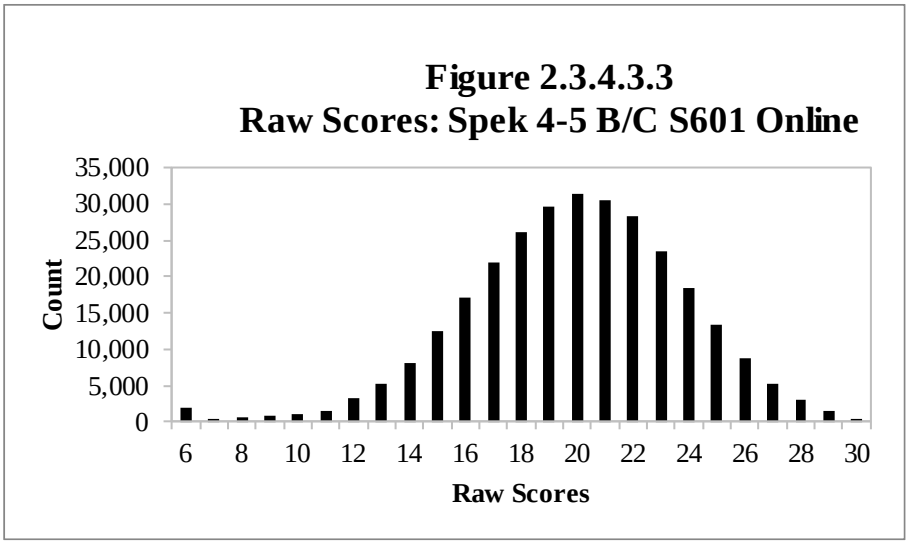


Table 2.3.4.3.3

Raw Score Descriptive Statistics: Spek 4-5 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	163,142	6	30	19.78	3.91
5	132,360	6	30	20.01	3.92
Total	295,502	6	30	19.88	3.92



2.3.4.4 Grades 6-8

Table 2.3.4.4.1
Raw Score Descriptive Statistics: Spek 6-8 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	2,916	0	6	4.71	1.97
7	5,269	0	6	4.72	2.01
8	8,313	0	6	4.81	1.95
Total	16,498	0	6	4.77	1.97

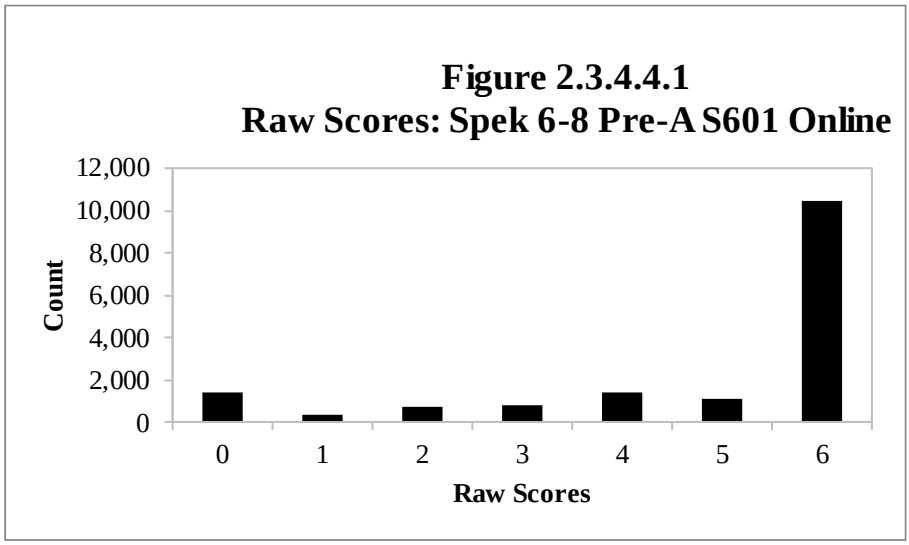


Table 2.3.4.4.2

Raw Score Descriptive Statistics: Spek 6-8 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	29,065	0	18	9.39	3.26
7	23,554	0	17	9.17	3.36
8	44,089	0	18	10.43	3.23
Total	96,708	0	18	9.81	3.32

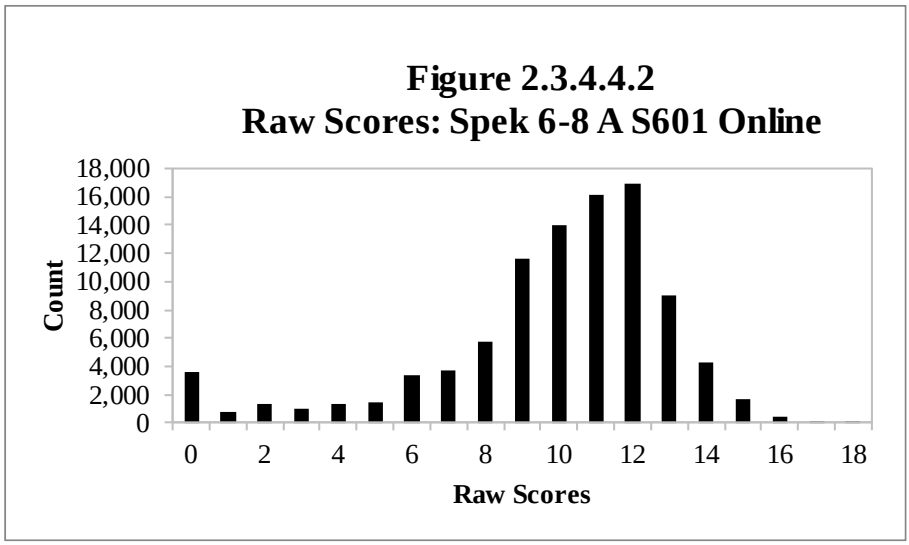
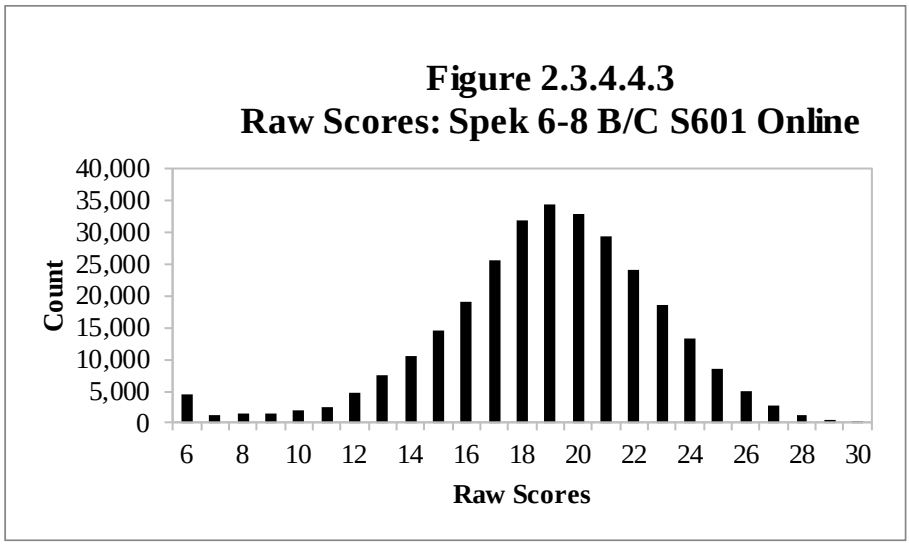


Table 2.3.4.4.3

Raw Score Descriptive Statistics: Spek 6-8 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	107,326	6	30	18.48	3.79
7	110,051	6	30	18.62	4.09
8	80,100	6	30	19.72	3.98
Total	297,477	6	30	18.87	3.99



2.3.4.5 Grades 9-12

Table 2.3.4.5.1
Raw Score Descriptive Statistics: Spek 9-12 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	8,881	0	6	4.53	2.01
10	9,217	0	6	4.93	1.83
11	8,195	0	6	5.08	1.77
12	7,614	0	6	5.14	1.81
Total	33,907	0	6	4.91	1.87

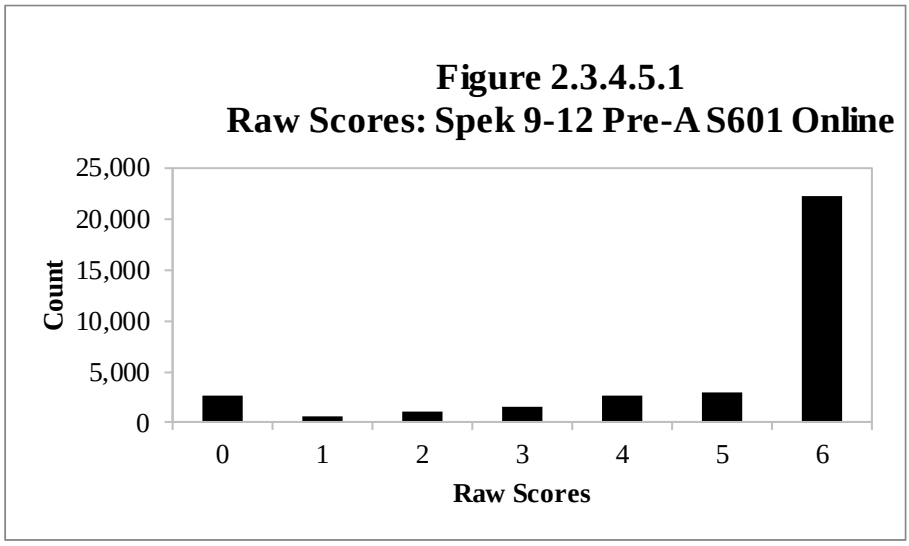


Table 2.3.4.5.2

Raw Score Descriptive Statistics: Spek 9-12 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	85,257	0	18	9.98	3.52
10	51,980	0	18	10.08	3.39
11	18,942	0	18	9.95	3.36
12	33,557	0	18	11.08	3.46
Total	189,736	0	18	10.20	3.48

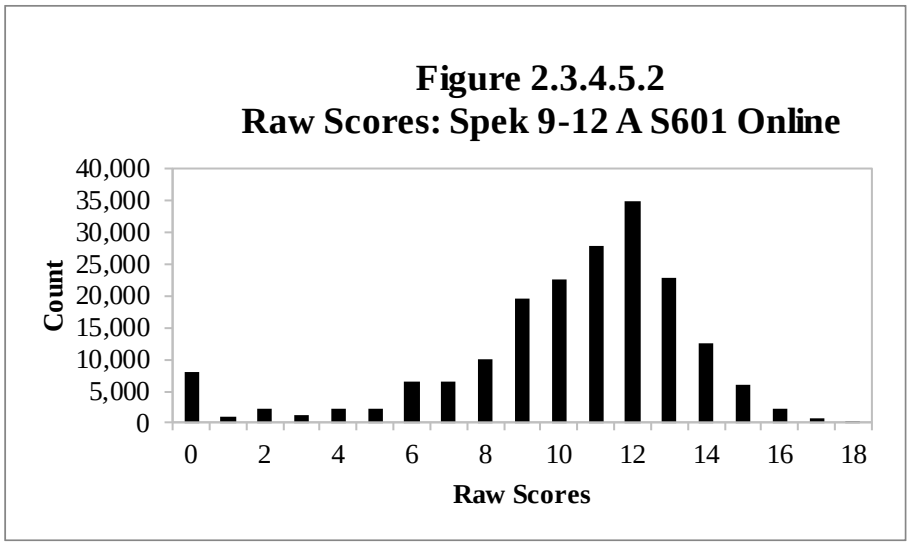
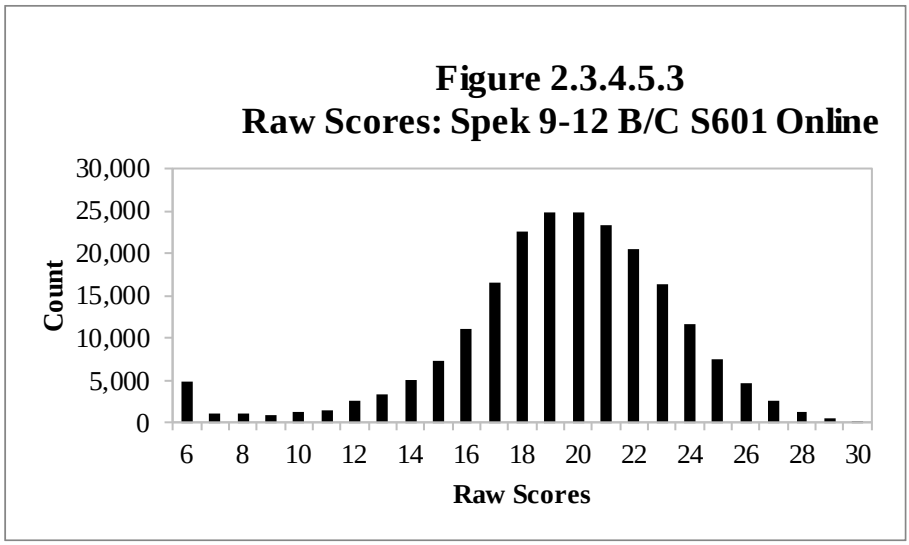


Table 2.3.4.5.3

Raw Score Descriptive Statistics: Spek 9-12 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	59,692	6	30	19.25	3.94
10	63,310	6	30	19.28	4.14
11	61,721	6	30	19.11	4.35
12	33,020	6	30	20.00	4.20
Total	217,743	6	30	19.33	4.17



2.4 Scale Score Distribution

Figures and tables in this section relate to the ACCESS for ELLs scale scores on each test form. For each test form, we converted raw scores to vertically equated scale scores. The scale score distributions are presented by grade-level cluster. Additionally, for Writing and Speaking, we present the distributions by grade-level cluster and tier.

For each test form, the figure shows the distribution of the scale scores. Scale scores are plotted on the horizontal axis.

For Listening and Reading, we grouped the scale scores into units of five scale score points (e.g., 100–104, 105–109, 110–114, etc.). It should be noted that the scale score distribution is presented by grade level cluster. Because the Listening and Reading domains are computer adaptive, students were routed by the engine into one of three different tier folders across stages, where the folders differ in difficulties. Therefore, in some plots below, it may appear that there is more than one set of data presented.

For Speaking and Writing, we plotted each individual scale score point for each test form. For figures that summarize both test forms in a cluster, we grouped scale scores into units of five scale score points.

It should be noted that Speaking Pre-A forms are designed for students at the very earliest stages of English language proficiency. Students routed to the Pre-A form have very low performances on Listening and Reading and are administered three Speaking tasks, each scored 0 to 2, for a total raw score range of 0 to 6. Tasks on the Pre-A form are by design very easy and intended to ensure beginning students are not discouraged. Therefore, large numbers of students can achieve all 6 points on this form as reflected in the Pre-A tables and figures below.

The number of students with scale scores falling into each range is plotted on the vertical axis.

The tables in this section show, by grade and by total for the grade-level cluster:

- The number of students in the analyses (count)
- The minimum observed scale score
- The maximum observed scale score

- The mean (average) scale score
- The standard deviation (std. dev.) of the scale score

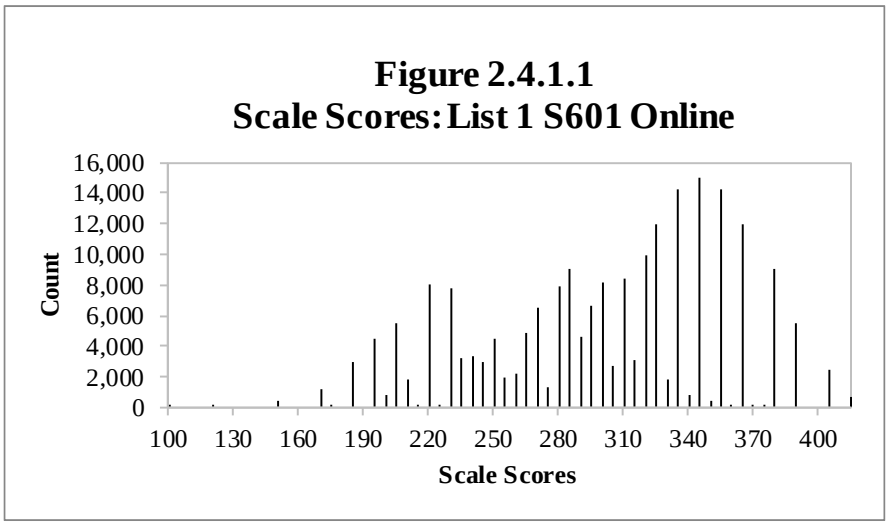
2.4.1 Listening

2.4.1.1 Grade 1

Table 2.4.1.1

Scale Score Descriptive Statistics: List 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	213,228	104	419	303.26	55.26
Total	213,228	104	419	303.26	55.26

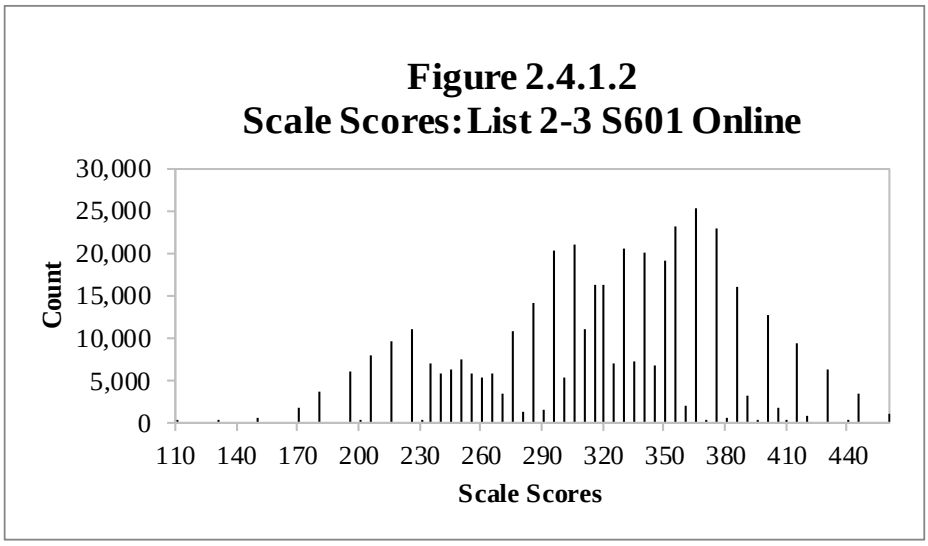


2.4.1.2 Grade 2-3

Table 2.4.1.2

Scale Score Descriptive Statistics: List 2-3 S601 Online

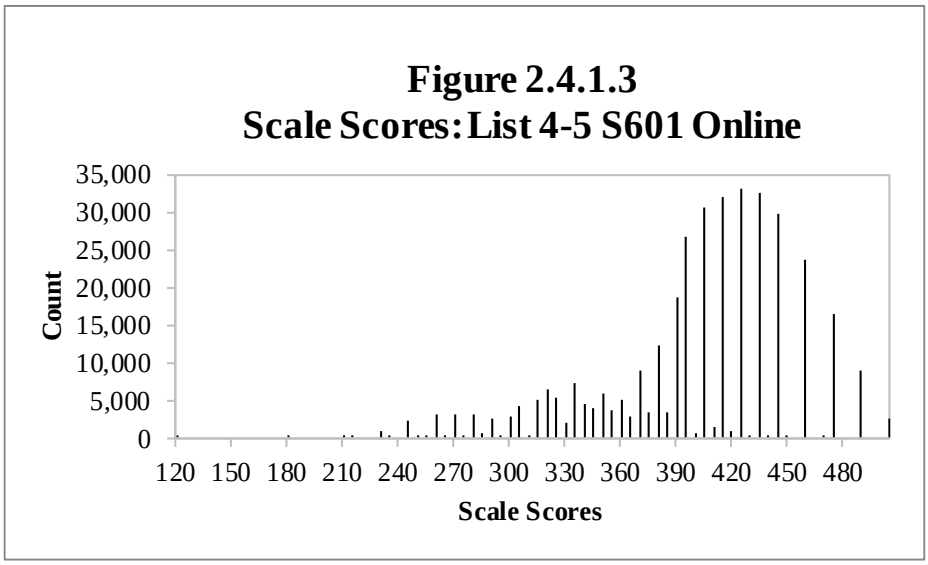
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	207,841	112	461	310.61	56.67
3	207,218	112	461	331.44	59.62
Total	415,059	112	461	321.01	59.08



2.4.1.3 Grade 4-5

Table 2.4.1.3
Scale Score Descriptive Statistics: List 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	199,010	120	508	401.00	51.20
5	165,820	120	508	405.42	56.19
Total	364,830	120	508	403.01	53.57

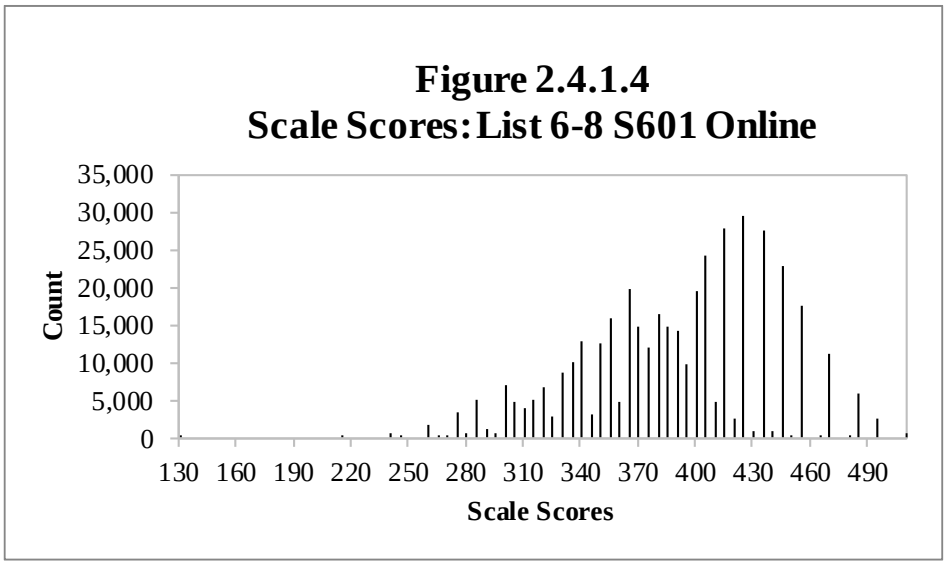


2.4.1.4 Grade 6-8

Table 2.4.1.4

Scale Score Descriptive Statistics: List 6-8 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	140,120	132	513	386.34	44.18
7	140,171	132	513	391.98	47.88
8	133,499	132	513	394.80	51.77
Total	413,790	132	513	390.98	48.11

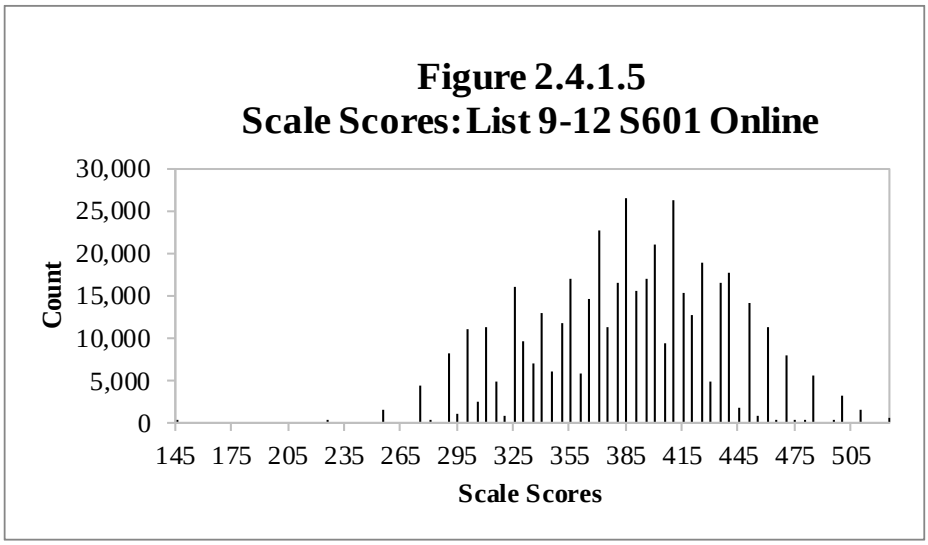


2.4.1.5 Grade 9-12

Table 2.4.1.5

Scale Score Descriptive Statistics: List 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	154,657	148	526	382.03	48.14
10	125,788	148	526	387.43	48.79
11	90,901	148	526	393.41	48.61
12	74,469	148	526	394.47	47.91
Total	445,815	148	526	387.95	48.64

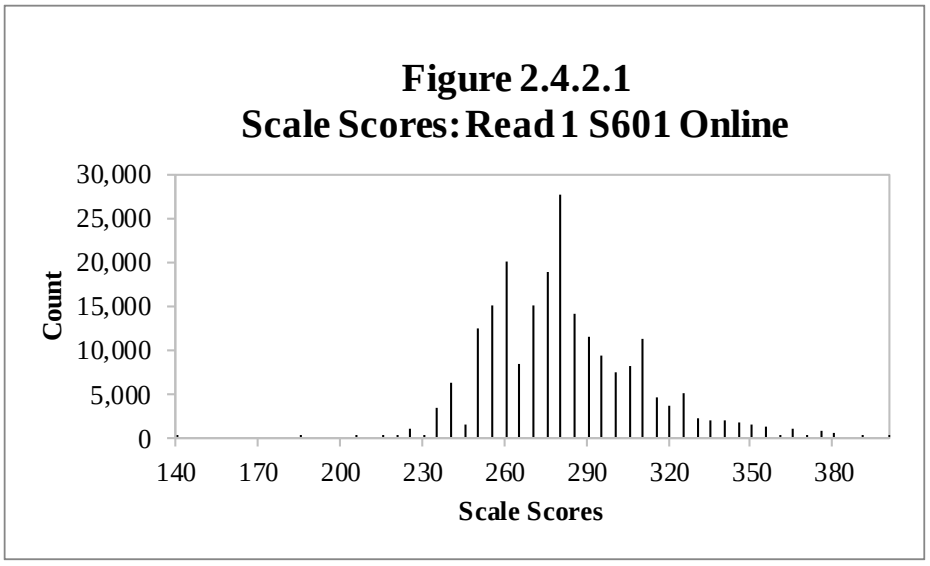


2.4.2 Reading

2.4.2.1 Grade 1

Table 2.4.2.1
Scale Score Descriptive Statistics: Read 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	218,896	141	402	283.68	27.89
Total	218,896	141	402	283.68	27.89

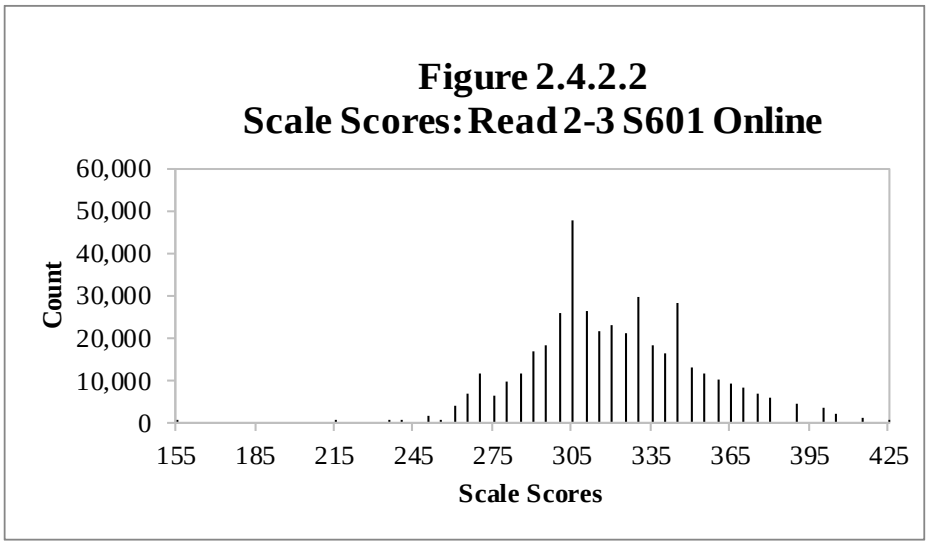


2.4.2.2 Grade 2-3

Table 2.4.2.2

Scale Score Descriptive Statistics: Read 2-3 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	213,322	158	427	317.44	26.80
3	210,005	158	427	328.44	33.77
Total	423,327	158	427	322.89	30.95

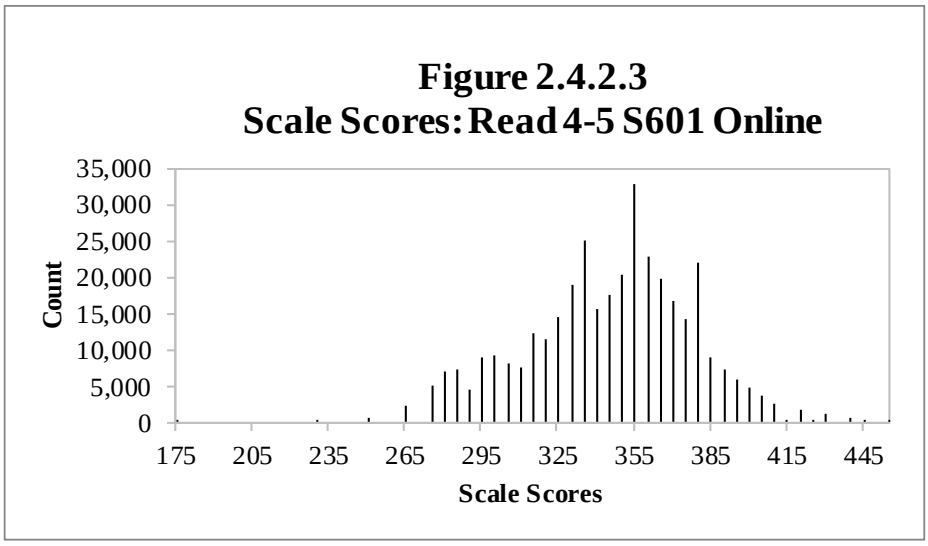


2.4.2.3 Grade 4-5

Table 2.4.2.3

Scale Score Descriptive Statistics: Read 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	198,383	175	458	345.48	32.33
5	164,720	175	458	348.77	34.10
Total	363,103	175	458	346.97	33.18

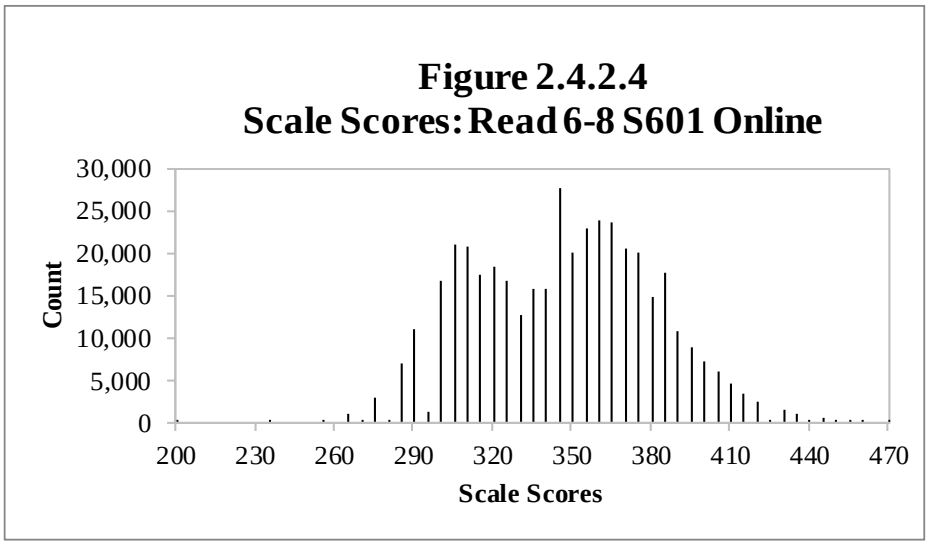


2.4.2.4 Grade 6-8

Table 2.4.2.4

Scale Score Descriptive Statistics: Read 6-8 S601 Online

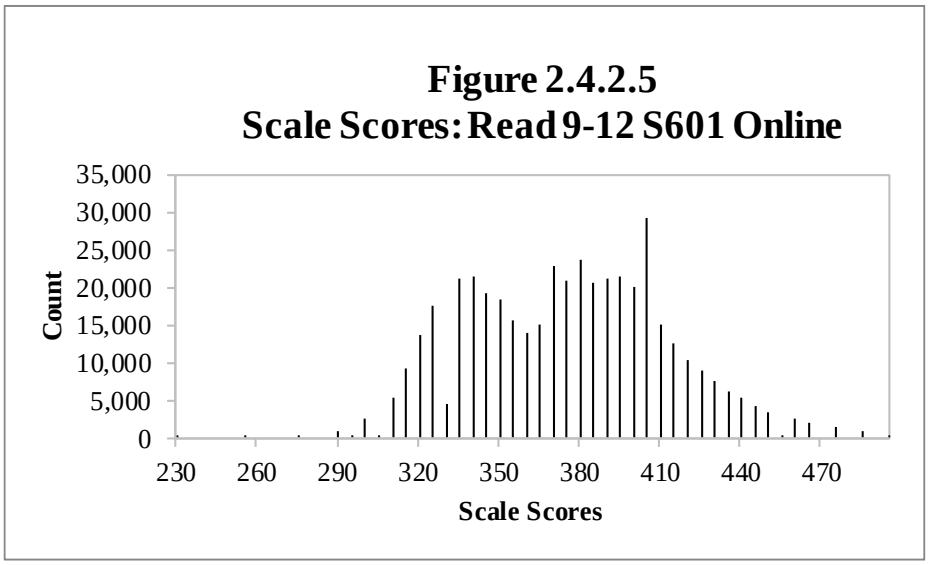
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	141,417	200	473	343.40	31.06
7	141,396	200	473	349.38	34.02
8	134,141	200	473	353.85	36.51
Total	416,954	200	473	348.79	34.16



2.4.2.5 Grade 9-12

Table 2.4.2.5
Scale Score Descriptive Statistics: Read 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	153,136	233	496	372.01	34.84
10	124,032	259	496	377.14	36.17
11	89,417	259	496	383.25	36.97
12	73,486	233	496	384.13	37.08
Total	440,071	233	496	377.76	36.37



2.4.3 Writing

2.4.3.1 Grade 1

Table 2.4.3.1.1

Scale Score Descriptive Statistics: Writ 1 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	205,653	111	362	231.64	46.50
Total	205,653	111	362	231.64	46.50

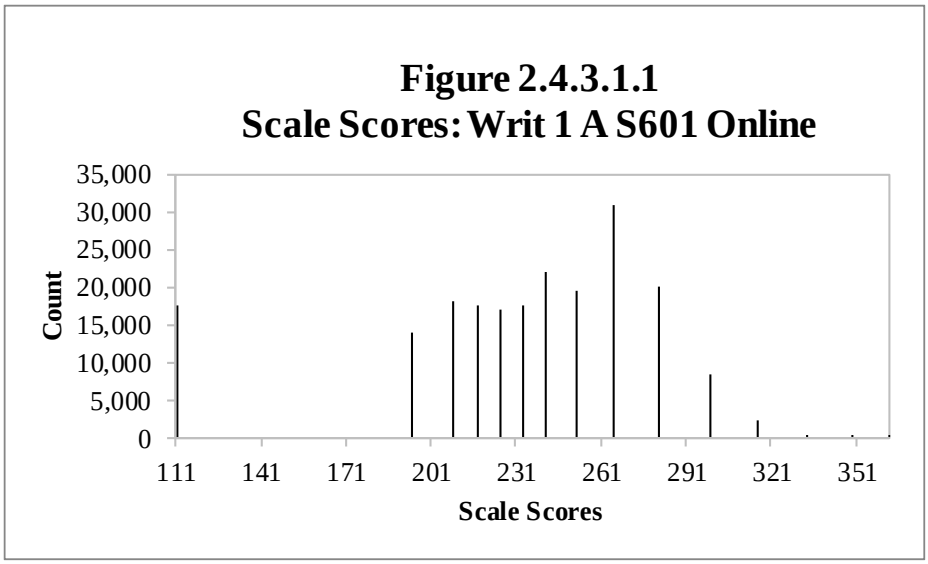


Table 2.4.3.1.2

Scale Score Descriptive Statistics: Writ 1 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	22,197	111	446	296.80	30.51
Total	22,197	111	446	296.80	30.51

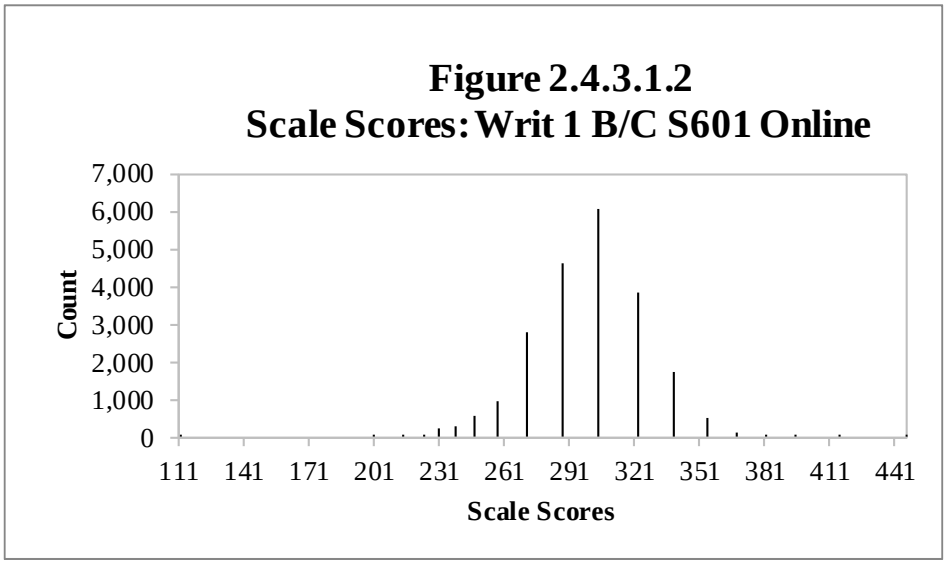
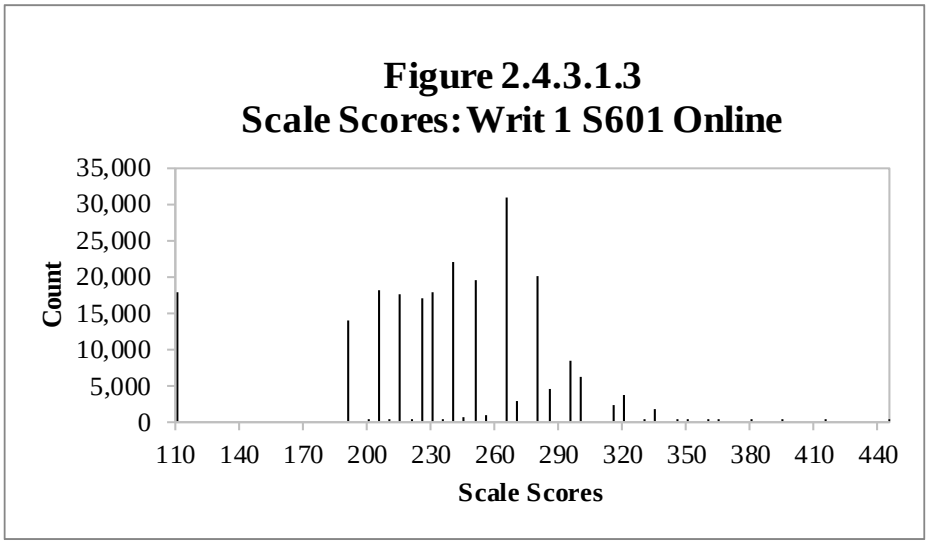


Table 2.4.3.1.3

Scale Score Descriptive Statistics: Writ 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	227,850	111	446	237.99	49.15
Total	227,850	111	446	237.99	49.15



2.4.3.2 Grade 2-3

Table 2.4.3.2.1

Scale Score Descriptive Statistics: Writ 2-3 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	74,987	133	385	242.06	45.35
3	62,236	133	385	252.01	45.08
Total	137,223	133	385	246.57	45.50

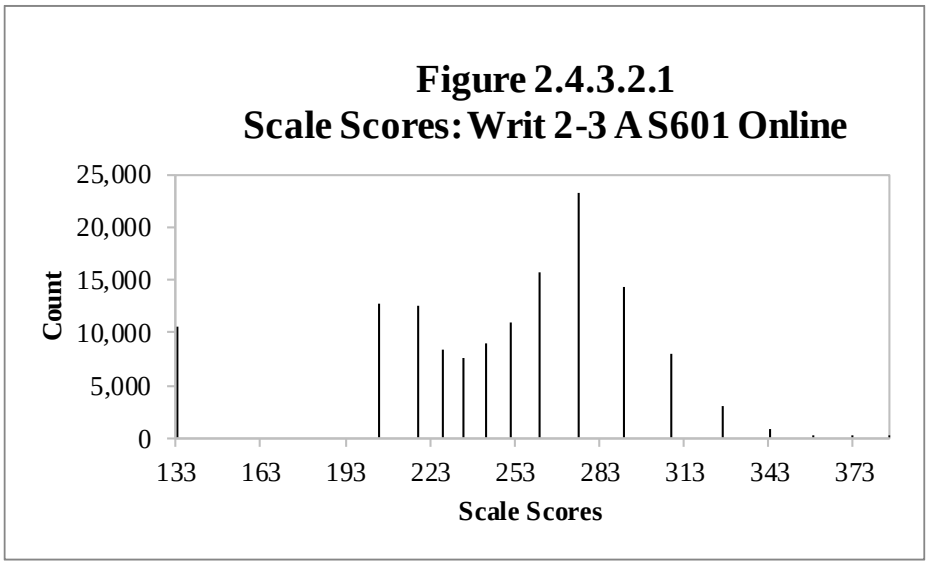


Table 2.4.3.2.2

Scale Score Descriptive Statistics: Writ 2-3 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	149,534	133	408	293.27	38.34
3	159,148	133	427	314.57	32.53
Total	308,682	133	427	304.25	37.03

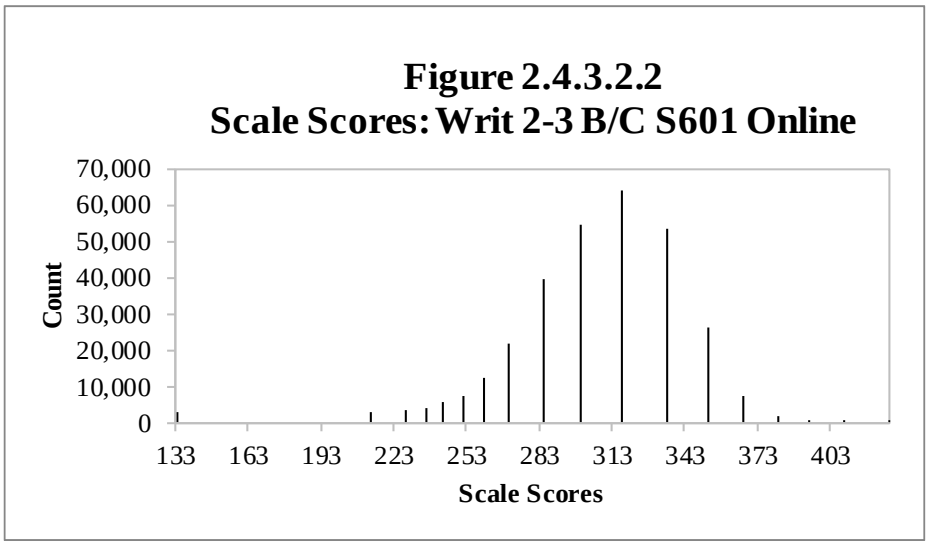
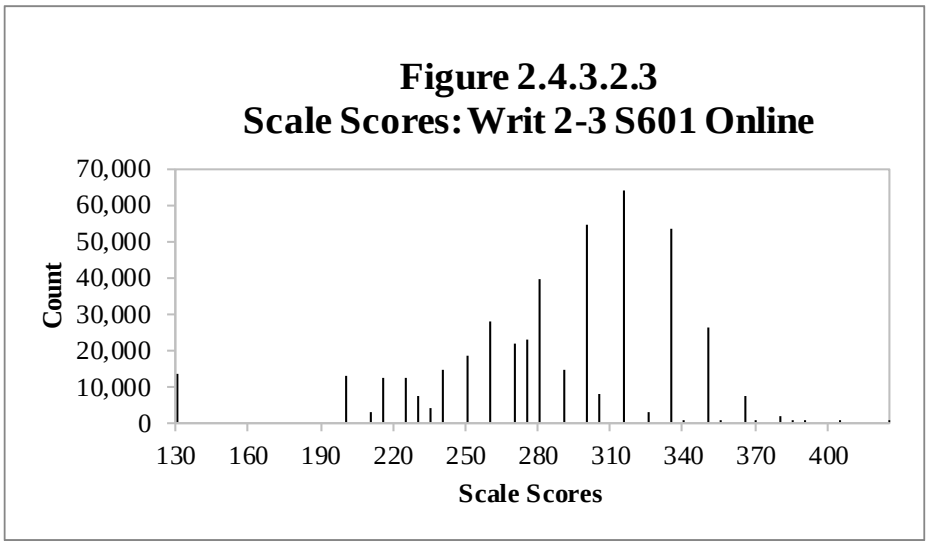


Table 2.4.3.2.3

Scale Score Descriptive Statistics: Writ 2-3 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	224,521	133	408	276.16	47.42
3	221,384	133	427	296.98	46.08
Total	445,905	133	427	286.50	47.90



2.4.3.3 Grade 4-5

Table 2.4.3.3.1

Scale Score Descriptive Statistics: Writ 4-5 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	43,240	155	416	257.04	51.40
5	43,207	155	416	268.67	50.24
Total	86,447	155	416	262.85	51.15

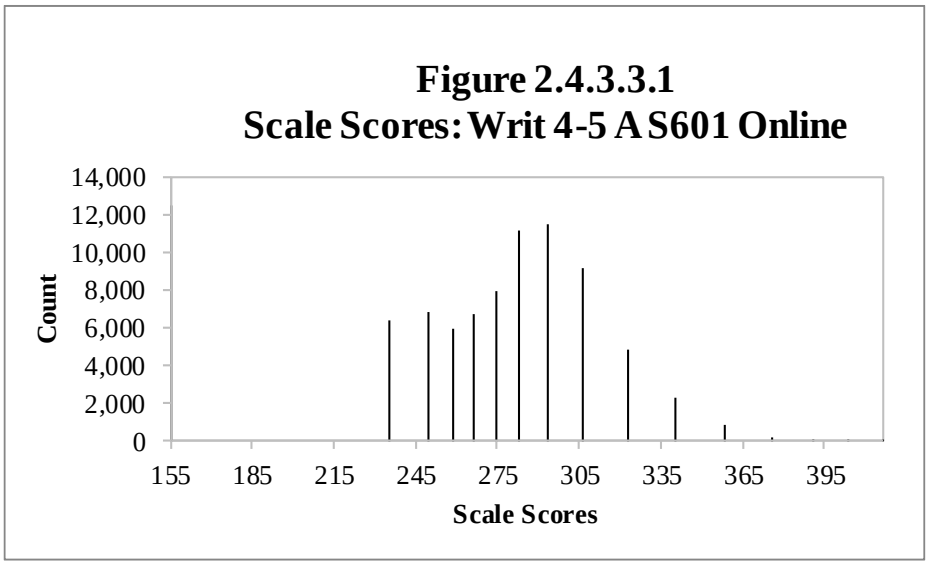


Table 2.4.3.3.2

Scale Score Descriptive Statistics: Writ 4-5 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	160,106	155	461	339.54	33.93
5	126,160	155	481	351.08	30.77
Total	286,266	155	481	344.63	33.08

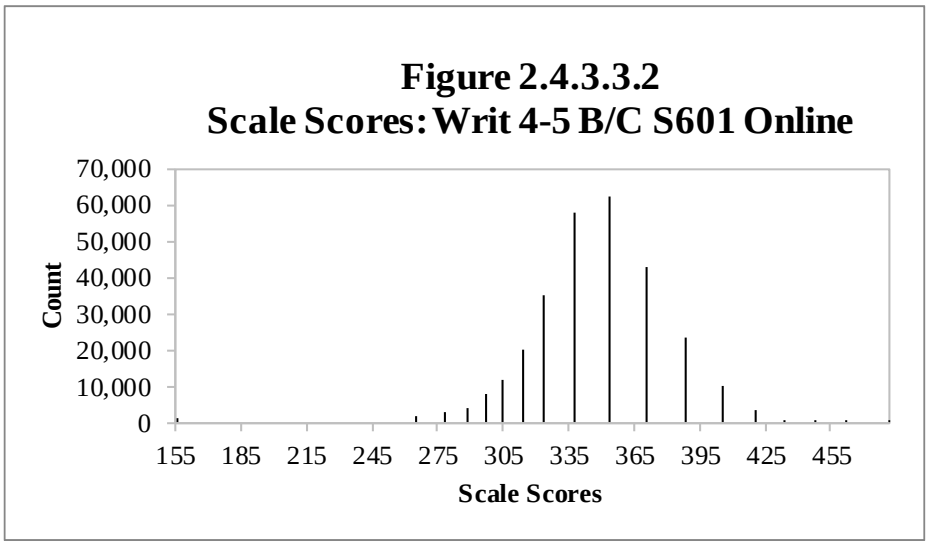
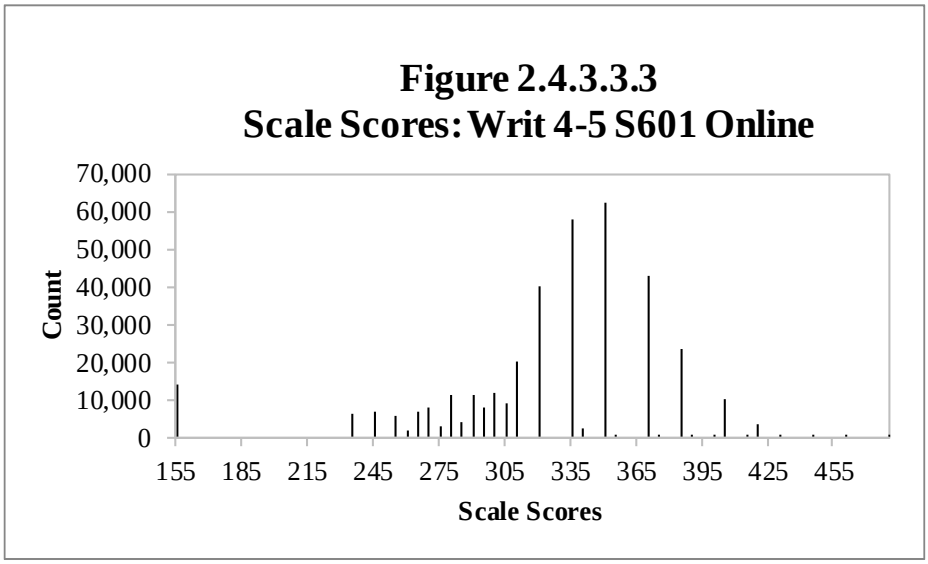


Table 2.4.3.3.3

Scale Score Descriptive Statistics: Writ 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	203,346	155	461	322.00	51.07
5	169,367	155	481	330.06	51.38
Total	372,713	155	481	325.66	51.37



2.4.3.4 Grade 6-8

Table 2.4.3.4.1
Scale Score Descriptive Statistics: Writ 6-8 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	49,195	188	421	278.41	38.90
7	60,604	188	472	288.47	39.70
8	63,844	188	472	295.24	40.52
Total	173,643	188	472	288.11	40.35

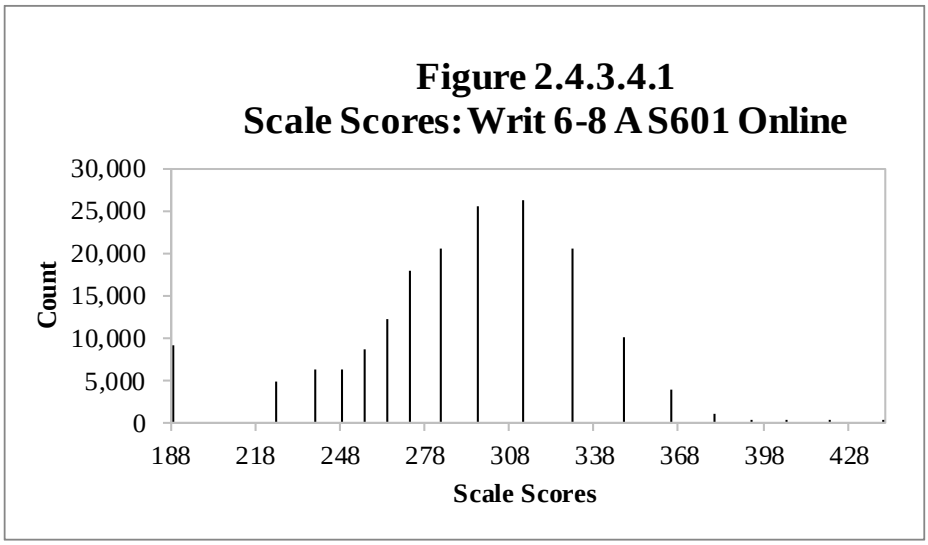


Table 2.4.3.4.2

Scale Score Descriptive Statistics: Writ 6-8 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	96,443	188	434	324.71	28.47
7	84,839	188	434	337.40	26.86
8	73,353	188	454	345.43	26.12
Total	254,635	188	454	334.91	28.59

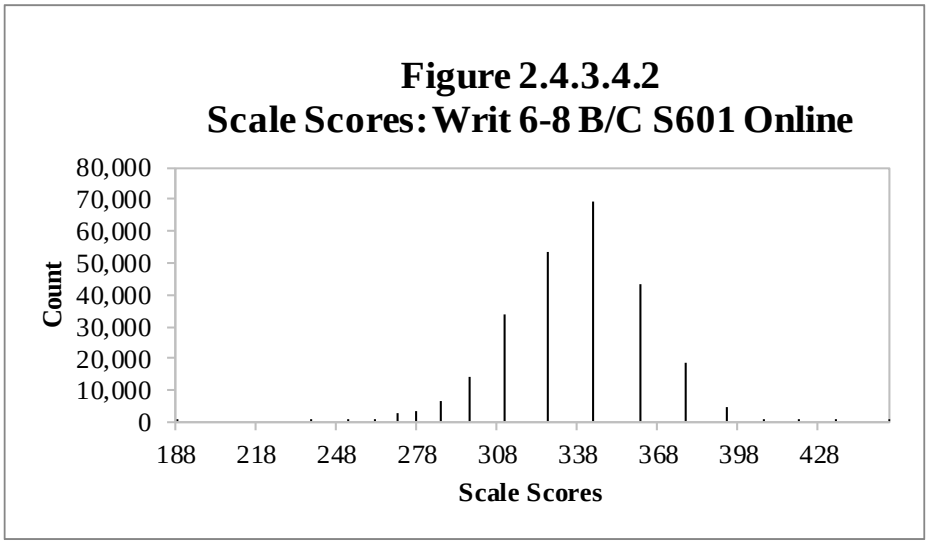
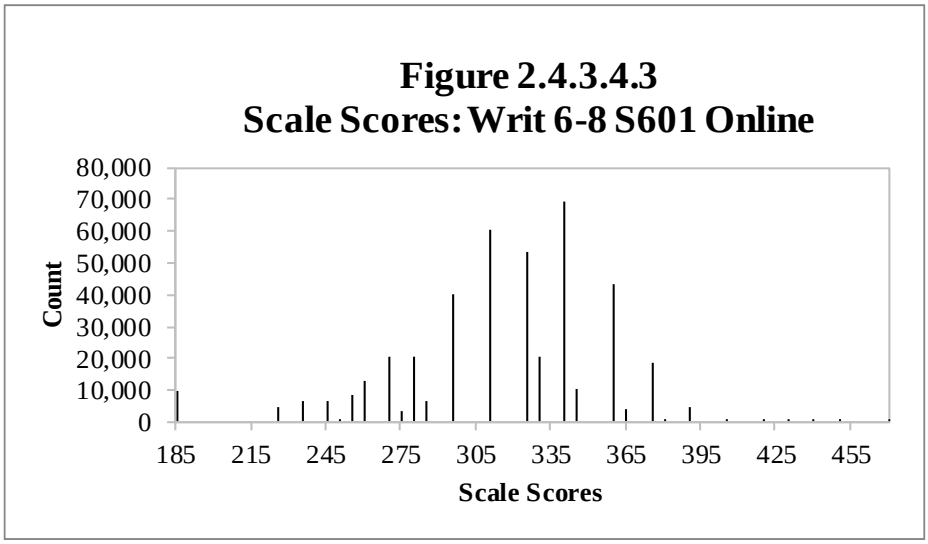


Table 2.4.3.4.3

Scale Score Descriptive Statistics: Writ 6-8 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	145,638	188	434	309.07	39.08
7	145,443	188	472	317.01	40.74
8	137,197	188	472	322.08	41.90
Total	428,278	188	472	315.94	40.92



2.4.3.5 Grade 9-12

Table 2.4.3.5.1
Scale Score Descriptive Statistics: Writ 9-12 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	66,714	232	477	310.57	40.97
10	49,466	232	508	316.32	37.15
11	32,094	232	457	323.17	36.42
12	24,237	232	457	325.10	36.96
Total	172,511	232	508	316.61	38.94

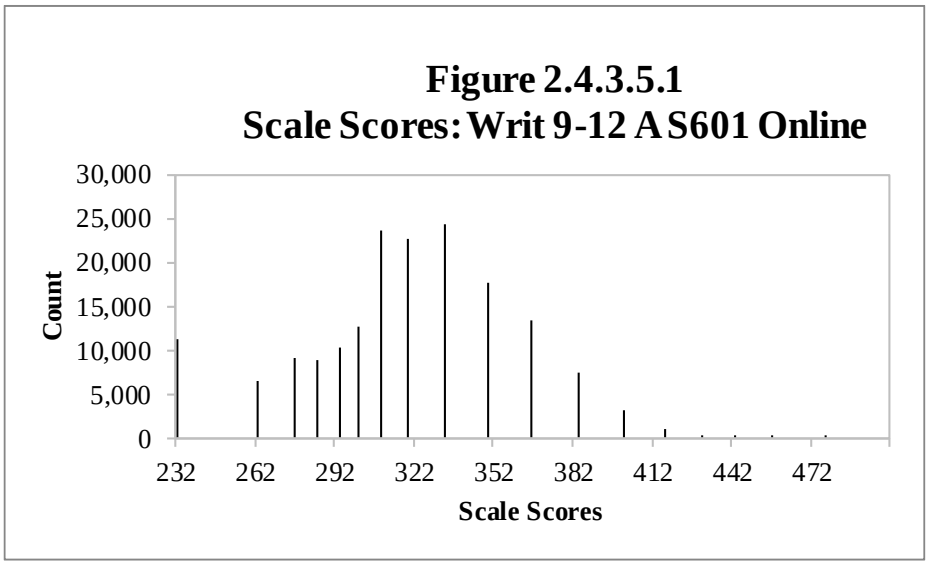


Table 2.4.3.5.2

Scale Score Descriptive Statistics: Writ 9-12 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	92,334	232	506	354.94	31.57
10	79,223	232	506	356.19	31.47
11	60,494	232	506	359.86	31.88
12	51,684	232	506	358.00	32.42
Total	283,735	232	506	356.90	31.82

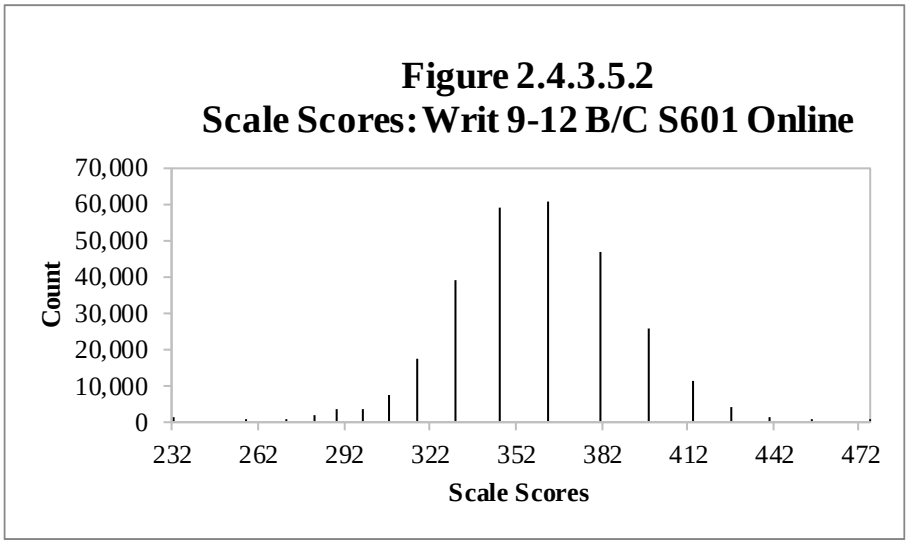
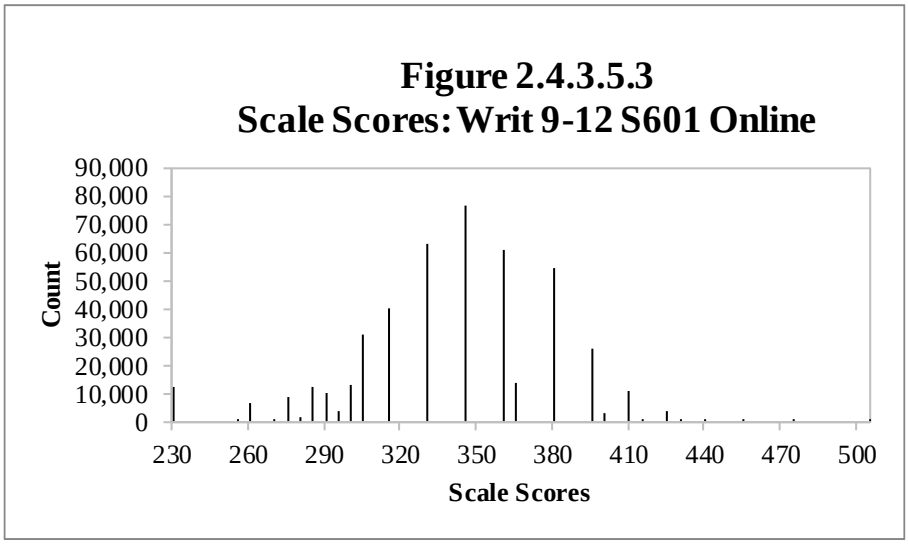


Table 2.4.3.5.3

Scale Score Descriptive Statistics: Writ 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	159,048	232	506	336.33	41.98
10	128,689	232	508	340.87	38.94
11	92,588	232	506	347.14	37.80
12	75,921	232	506	347.50	37.24
Total	456,246	232	508	341.66	39.81



2.4.4 Speaking

2.4.4.1 Grade 1

Table 2.4.4.1.1

Scale Score Descriptive Statistics: Spek 1 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	13,774	106	167	147.89	24.23
Total	13,774	106	167	147.89	24.23

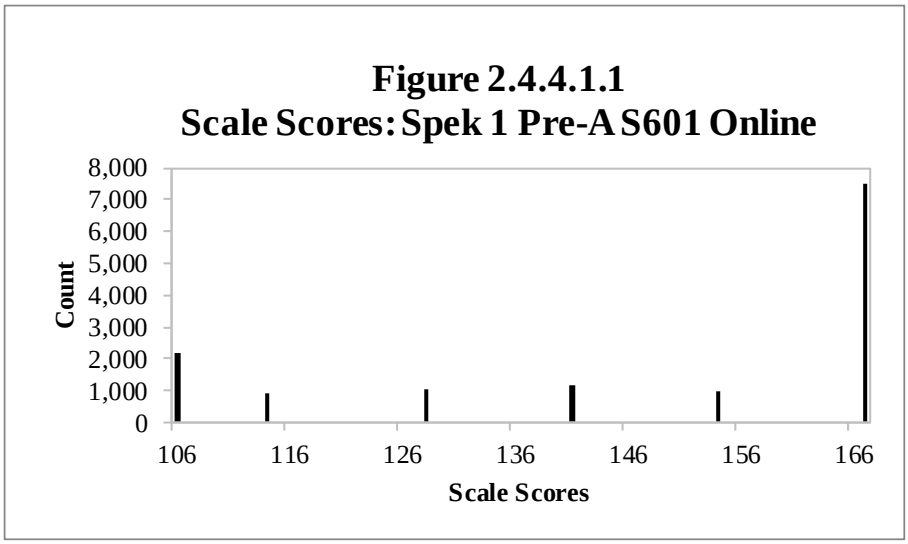


Table 2.4.4.1.2

Scale Score Descriptive Statistics: Spek 1 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	101,779	106	381	224.93	52.83
Total	101,779	106	381	224.93	52.83

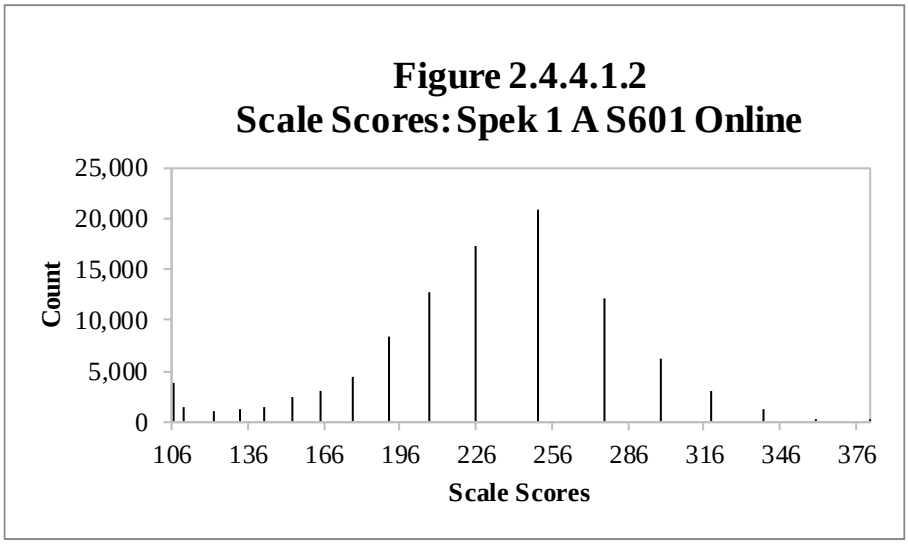


Table 2.4.4.1.3

Scale Score Descriptive Statistics: Spek 1 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	95,004	106	403	269.71	44.08
Total	95,004	106	403	269.71	44.08

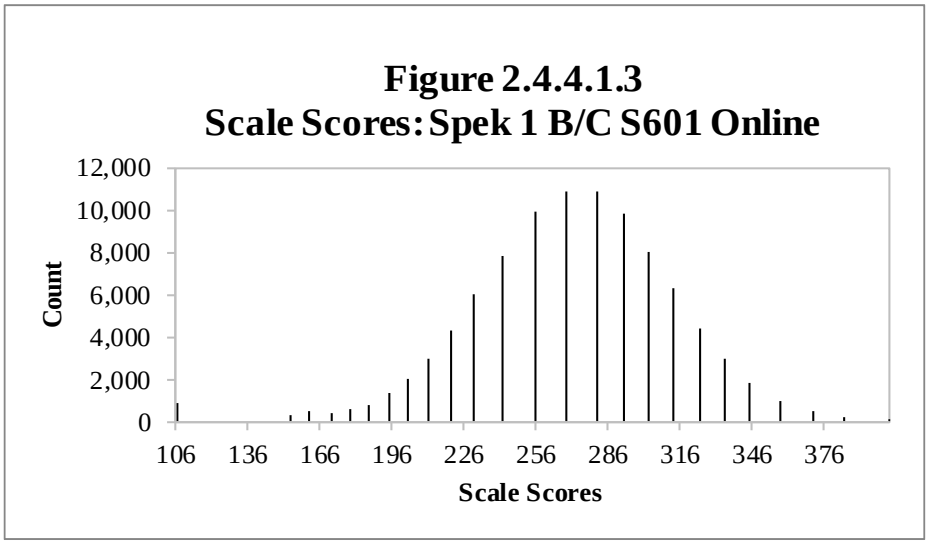
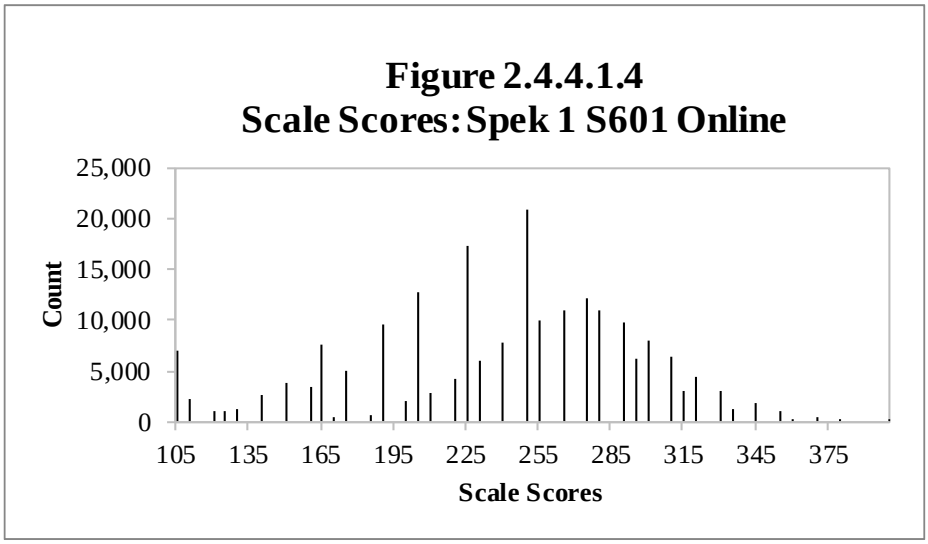


Table 2.4.4.1.4

Scale Score Descriptive Statistics: Spek 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	210,557	106	403	240.09	57.68
Total	210,557	106	403	240.09	57.68



2.4.4.2 Grade 2-3

Table 2.4.4.2.1

Scale Score Descriptive Statistics: Spek 2-3 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	6,222	118	168	153.06	20.51
3	14,627	118	168	154.57	19.90
Total	20,849	118	168	154.12	20.10

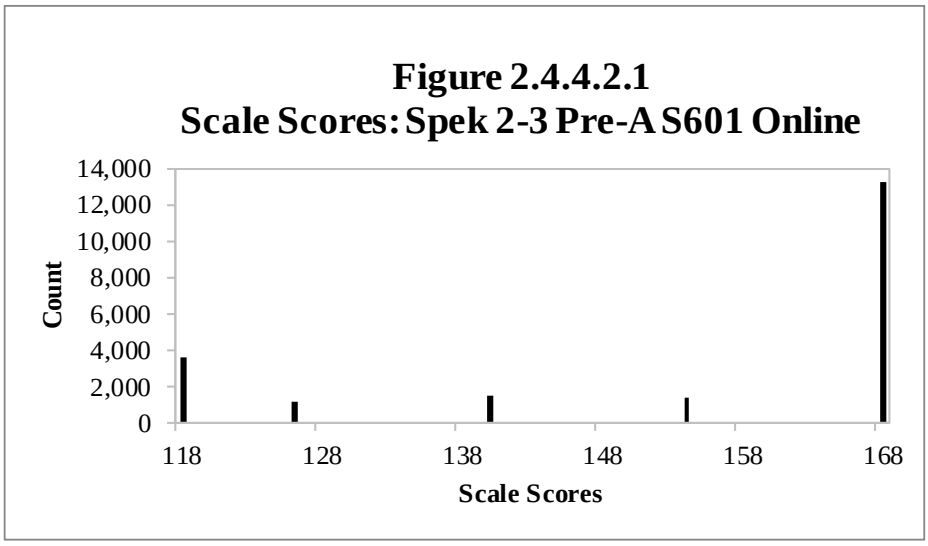


Table 2.4.4.2.2

Scale Score Descriptive Statistics: Spek 2-3 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	61,228	118	395	231.82	57.02
3	60,840	118	395	256.70	56.16
Total	122,068	118	395	244.22	57.94

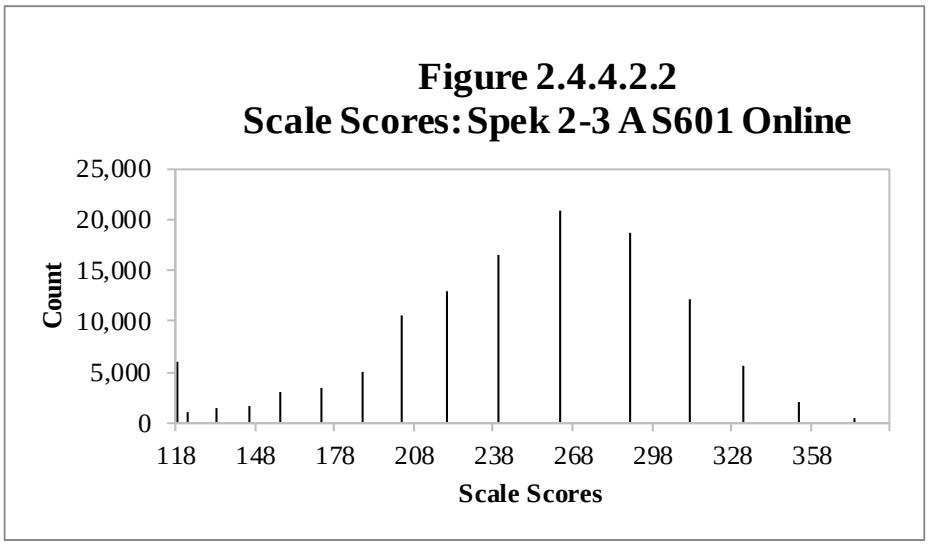


Table 2.4.4.2.3

Scale Score Descriptive Statistics: Spek 2-3 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	139,524	118	425	282.39	45.58
3	130,550	118	425	303.41	43.33
Total	270,074	118	425	292.55	45.73

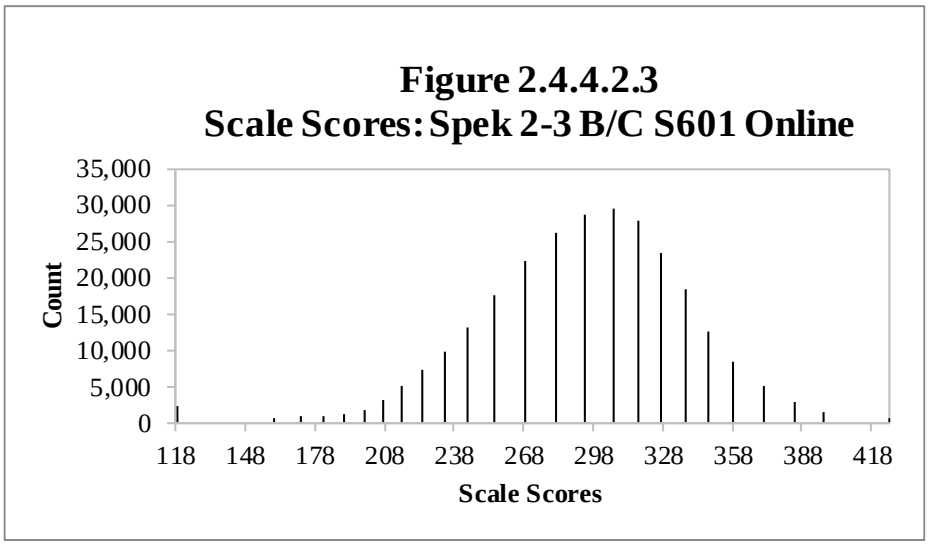
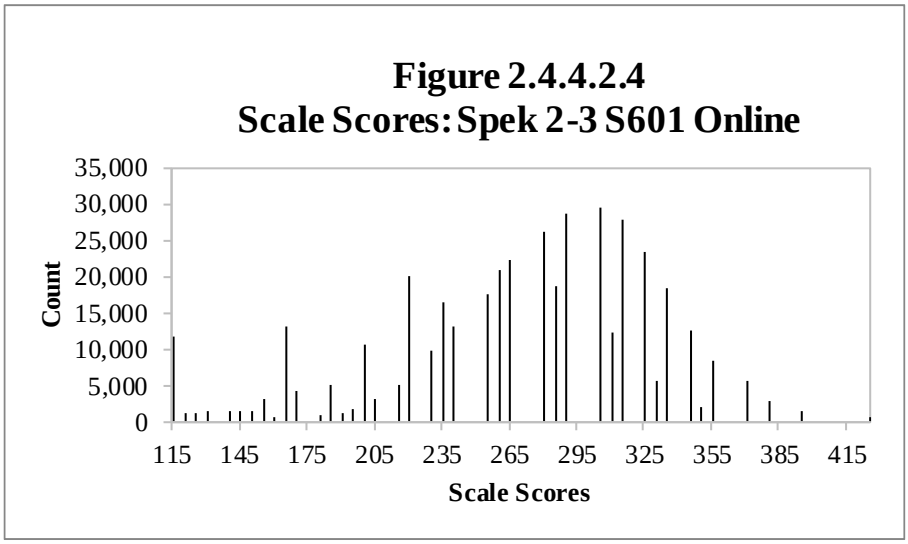


Table 2.4.4.2.4

Scale Score Descriptive Statistics: Spek 2-3 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	206,974	118	425	263.54	57.26
3	206,017	118	425	279.05	61.42
Total	412,991	118	425	271.28	59.88



2.4.4.3 Grade 4-5

Table 2.4.4.3.1

Scale Score Descriptive Statistics: Spek 4-5 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	2,426	130	191	170.38	23.91
5	6,115	130	191	173.34	23.05
Total	8,541	130	191	172.50	23.33

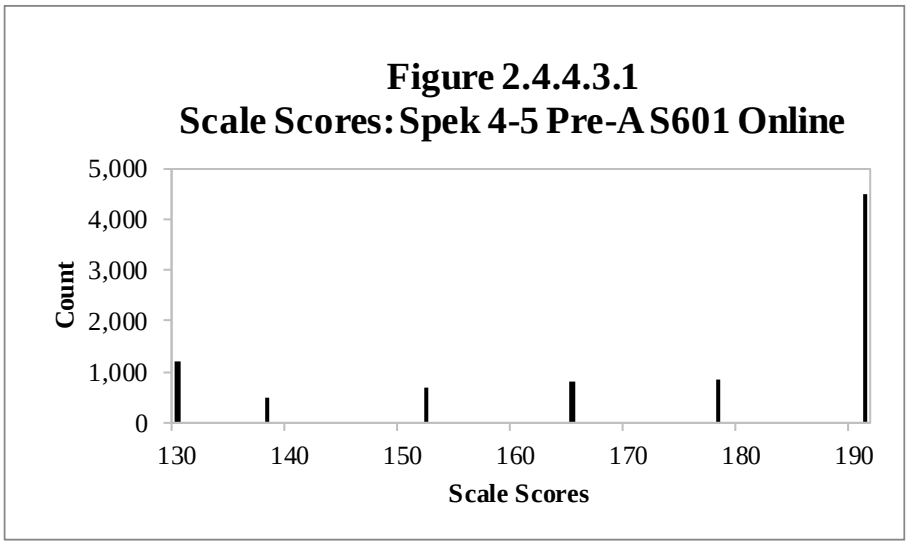


Table 2.4.4.3.2

Scale Score Descriptive Statistics: Spek 4-5 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	32,192	130	429	255.06	59.67
5	25,549	130	429	260.34	59.44
Total	57,741	130	429	257.39	59.62

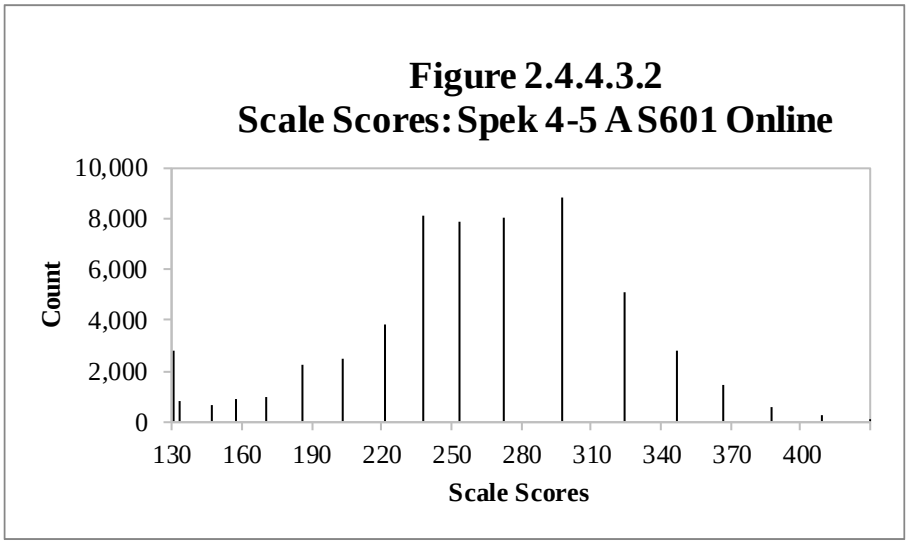


Table 2.4.4.3.3

Scale Score Descriptive Statistics: Spek 4-5 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	163,142	130	448	330.62	45.61
5	132,360	130	448	333.18	45.76
Total	295,502	130	448	331.77	45.69

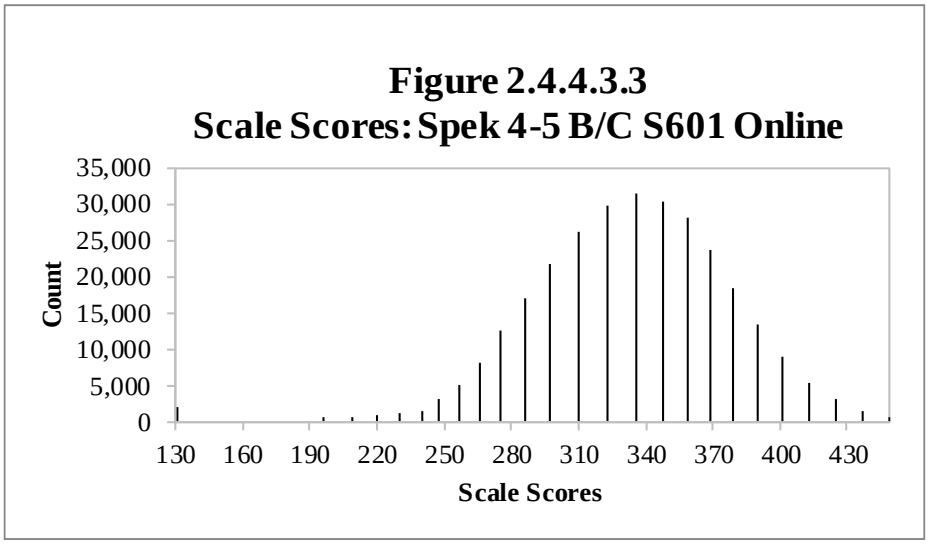
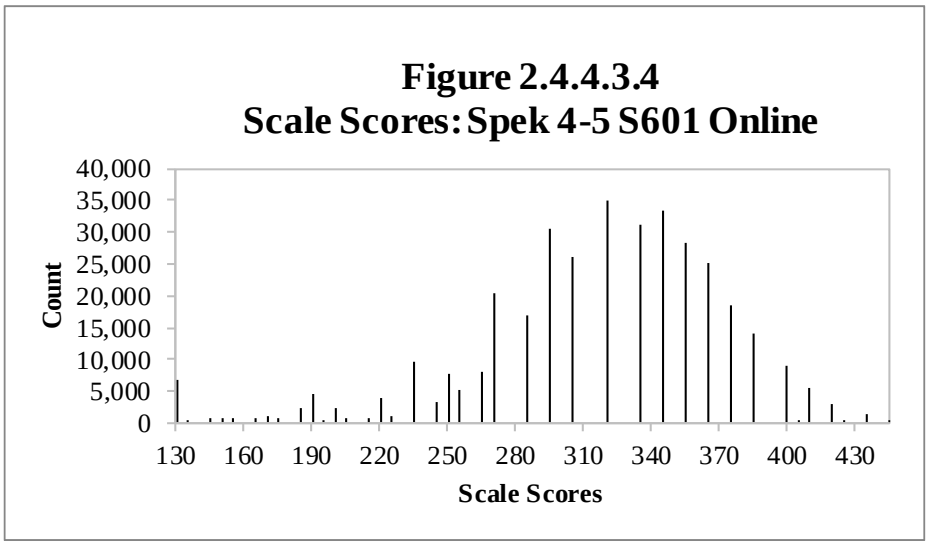


Table 2.4.4.3.4

Scale Score Descriptive Statistics: Spek 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	197,760	130	448	316.35	57.82
5	164,024	130	448	315.88	61.15
Total	361,784	130	448	316.14	59.35



2.4.4.4 Grade 6-8

Table 2.4.4.4.1

Scale Score Descriptive Statistics: Spek 6-8 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	2,916	148	220	202.97	25.29
7	5,269	148	220	203.22	25.62
8	8,313	148	220	204.35	24.93
Total	16,498	148	220	203.74	25.22

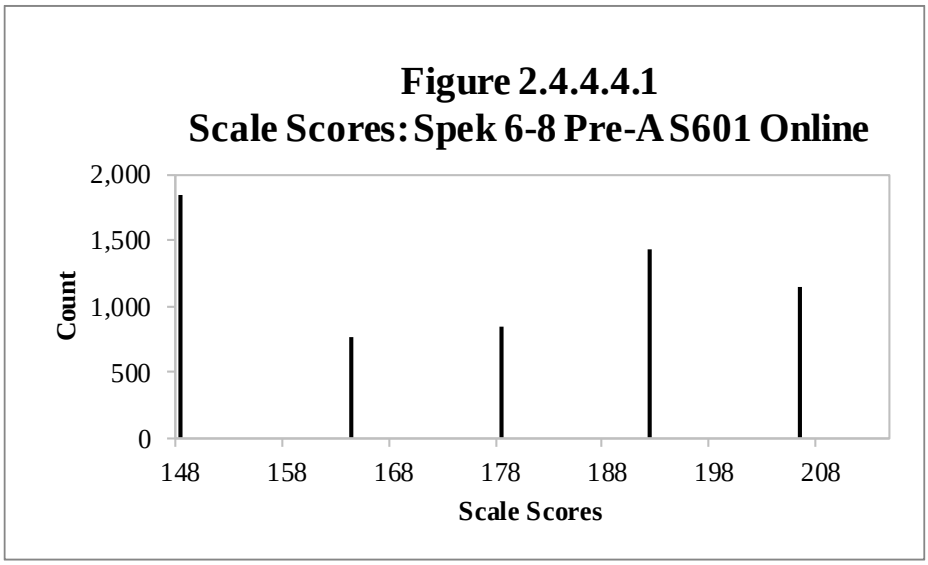


Table 2.4.4.4.2

Scale Score Descriptive Statistics: Spek 6-8 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	29,065	148	441	263.10	52.05
7	23,554	148	420	259.57	52.54
8	44,089	148	441	283.11	54.74
Total	96,708	148	441	271.36	54.50

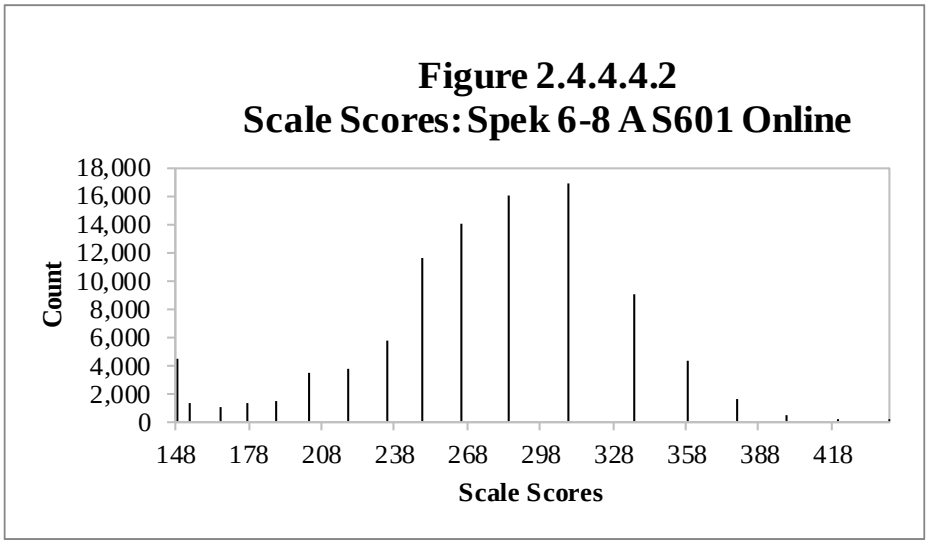


Table 2.4.4.4.3

Scale Score Descriptive Statistics: Spek 6-8 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	107,326	148	463	324.25	44.67
7	110,051	148	463	325.83	48.31
8	80,100	148	463	338.49	47.12
Total	297,477	148	463	328.67	47.08

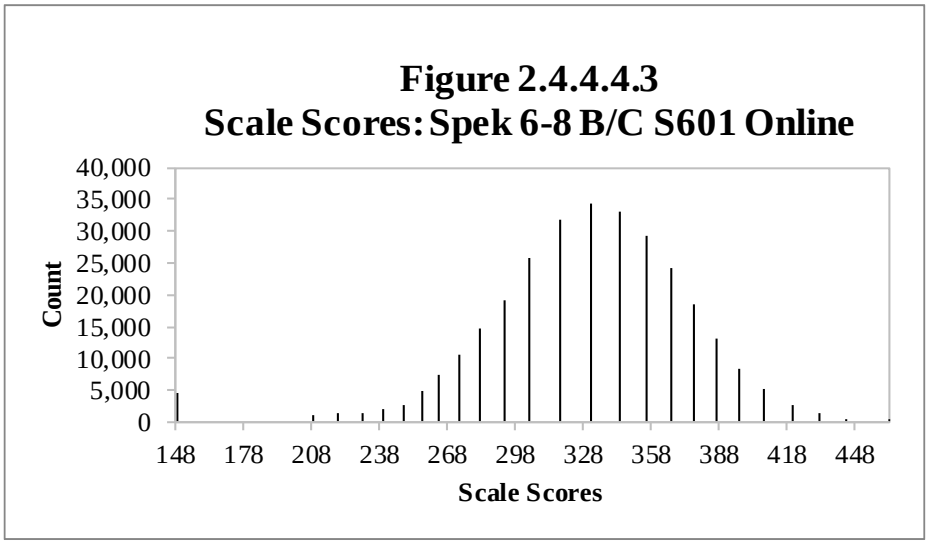
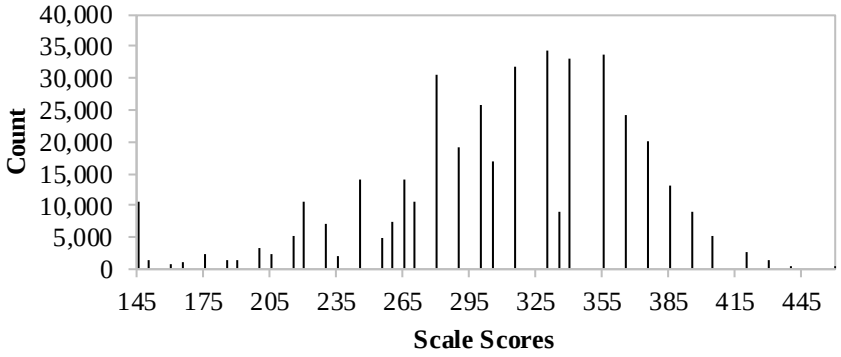


Table 2.4.4.4.4

Scale Score Descriptive Statistics: Spek 6-8 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	139,307	148	463	308.96	54.50
7	138,874	148	463	309.94	58.35
8	132,502	148	463	311.64	61.70
Total	410,683	148	463	310.16	58.21

Figure 2.4.4.4.4
Scale Scores: Spek 6-8 S601 Online



2.4.4.5 Grade 9-12

Table 2.4.4.5.1

Scale Score Descriptive Statistics: Spek 9-12 Pre-A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	8,881	172	222	206.08	19.70
10	9,217	172	222	210.50	18.01
11	8,195	172	222	212.31	17.18
12	7,614	172	222	213.32	17.02
Total	33,907	172	222	210.41	18.27

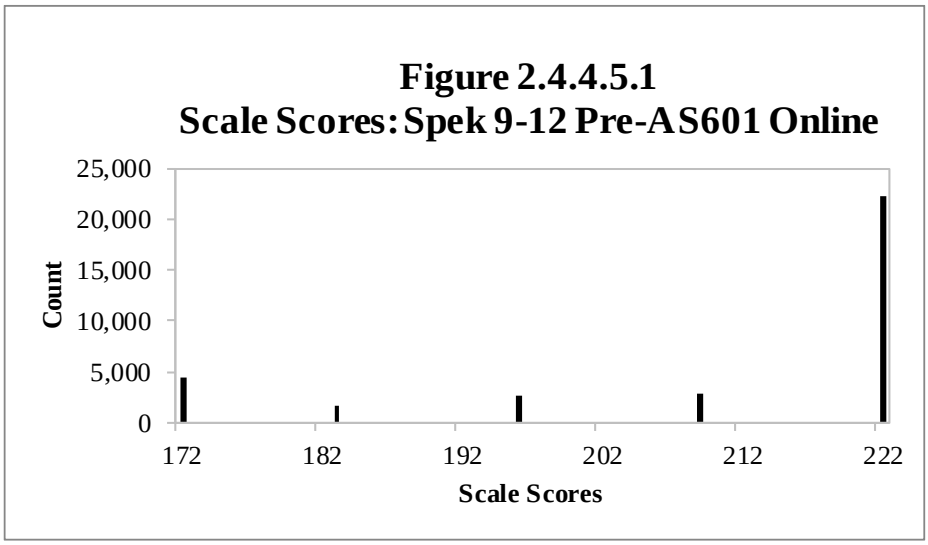


Table 2.4.4.5.2

Scale Score Descriptive Statistics: Spek 9-12 A S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	85,257	172	442	278.88	55.00
10	51,980	172	442	279.73	52.98
11	18,942	172	442	277.21	51.32
12	33,557	172	442	299.96	56.15
Total	189,736	172	442	282.68	54.89

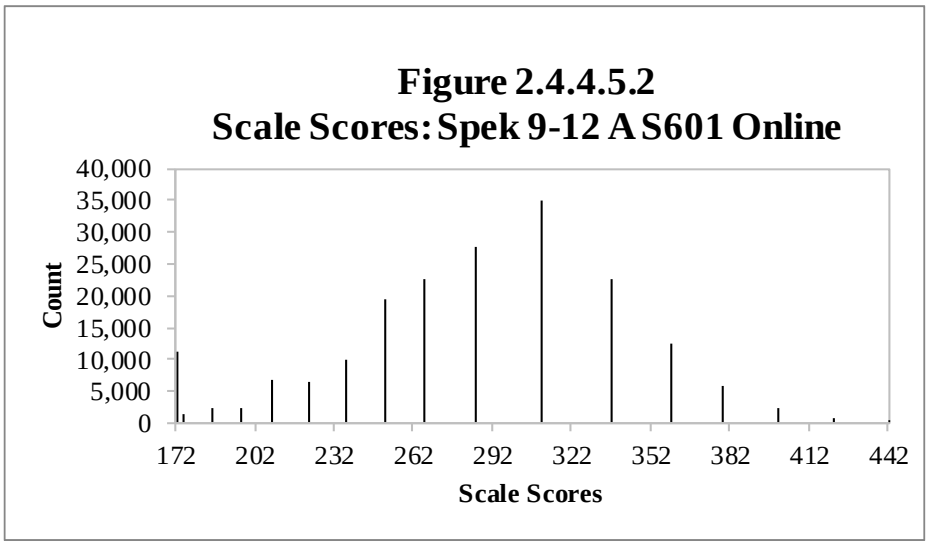


Table 2.4.4.5.3

Scale Score Descriptive Statistics: Spek 9-12 B/C S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	59,692	172	476	336.00	45.40
10	63,310	172	476	336.27	47.71
11	61,721	172	476	334.35	50.15
12	33,020	172	476	344.62	48.41
Total	217,743	172	476	336.92	48.02

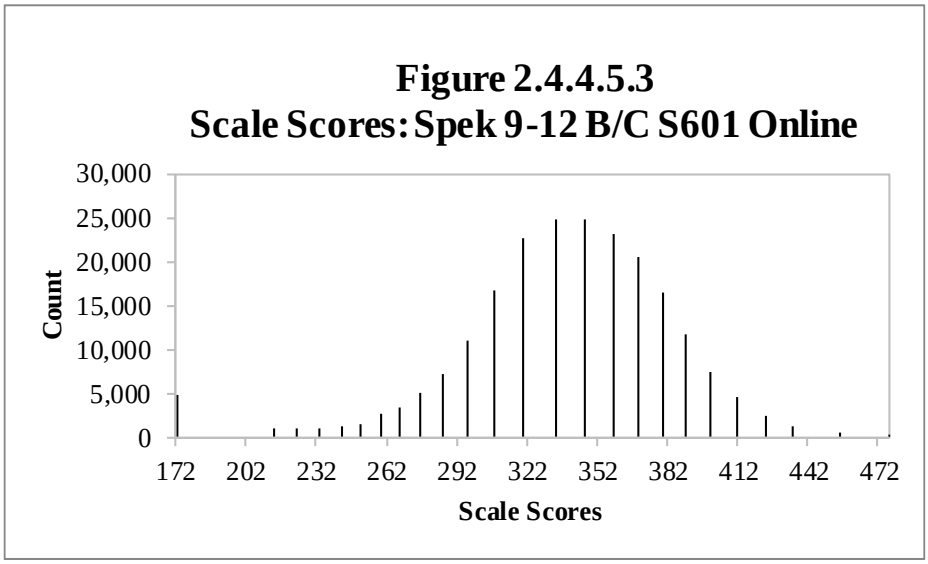
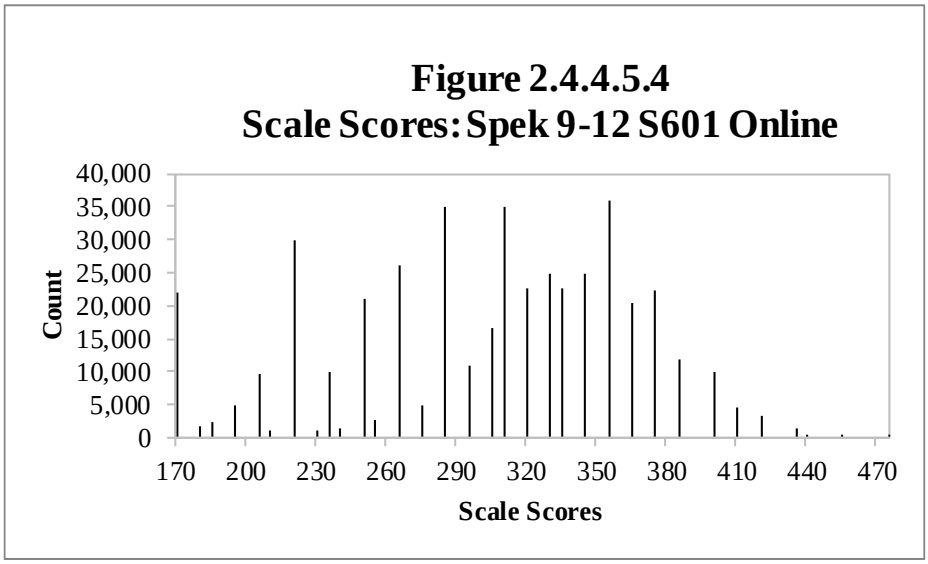


Table 2.4.4.5.4

Scale Score Descriptive Statistics: Spek 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	153,830	172	476	296.84	61.22
10	124,507	172	476	303.35	61.44
11	88,858	172	476	310.91	62.09
12	74,191	172	476	310.95	63.53
Total	441,386	172	476	303.88	62.14



2.5 Proficiency Level Distributions

The figures and tables in this section provide information about the proficiency level distributions of the students who took each test form based on their performance by grade-level cluster. For Writing and Speaking, we also present that information by grade-level cluster and tier.

In the tables presented in this section, each row shows, by grade and by total for the grade-level cluster:

- The WIDA proficiency level designation (1–6)
- The number of students (count) whose performance on the test form placed them into that proficiency level in the tested domain
- The percentage of students, out of the total number of students taking the form, who were placed into that proficiency level in the tested domain

In the figure, the horizontal axis shows the six WIDA proficiency levels. The vertical axis shows the percentage of students. Each bar shows the percentage of students who were placed into each proficiency level in the domain on this test form.

Note that WIDA intends for students who are just beginning to learn English to take the Speaking Pre-A tier; therefore, WIDA does not expect students assigned to this tier to show proficiency above PL 1.

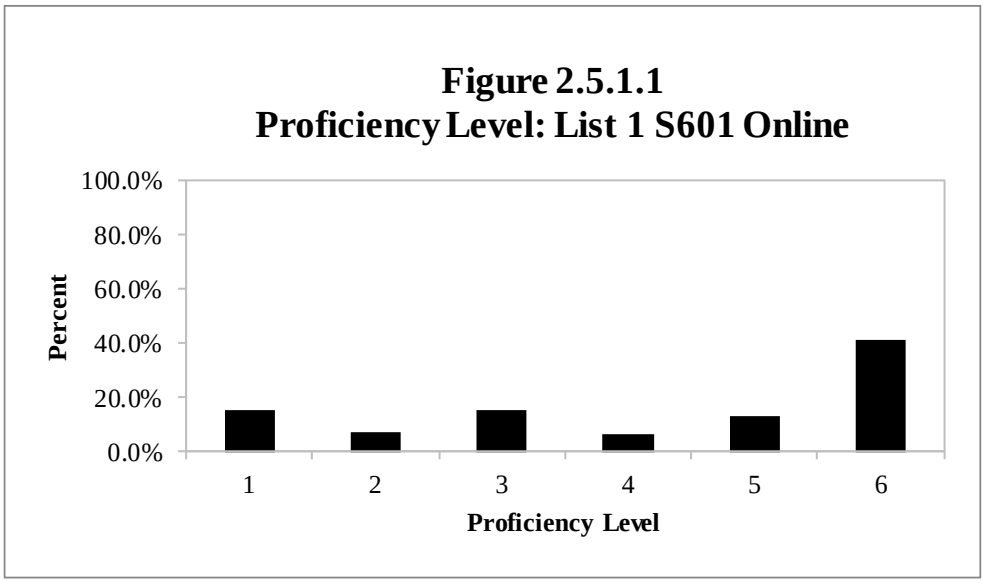
2.5.1 Listening

2.5.1.1 Grade 1

Table 2.5.1.1

Proficiency Level Distribution: List 1 S601 Online

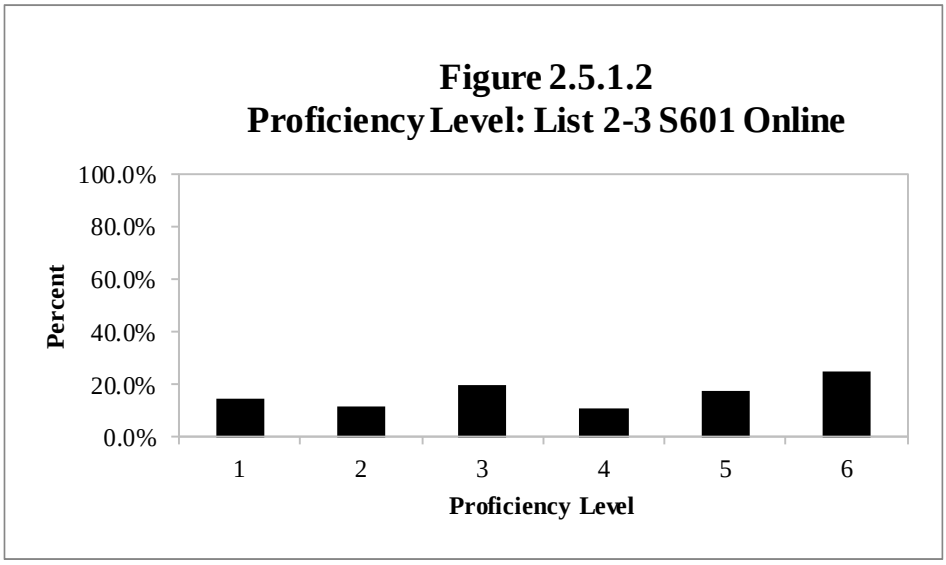
Level	Grade 1		Total	
	Count	**Percent**	**Count**	**Percent**
1	33,262	15.60%	33,262	15.60%
2	15,847	7.43%	15,847	7.43%
3	32,747	15.36%	32,747	15.36%
4	14,135	6.63%	14,135	6.63%
5	29,086	13.64%	29,086	13.64%
6	88,151	41.34%	88,151	41.34%
Total	213,228	100.00%	213,228	100.00%



2.5.1.2 Grade 2-3

Table 2.5.1.2
Proficiency Level Distribution: List 2-3 S601 Online

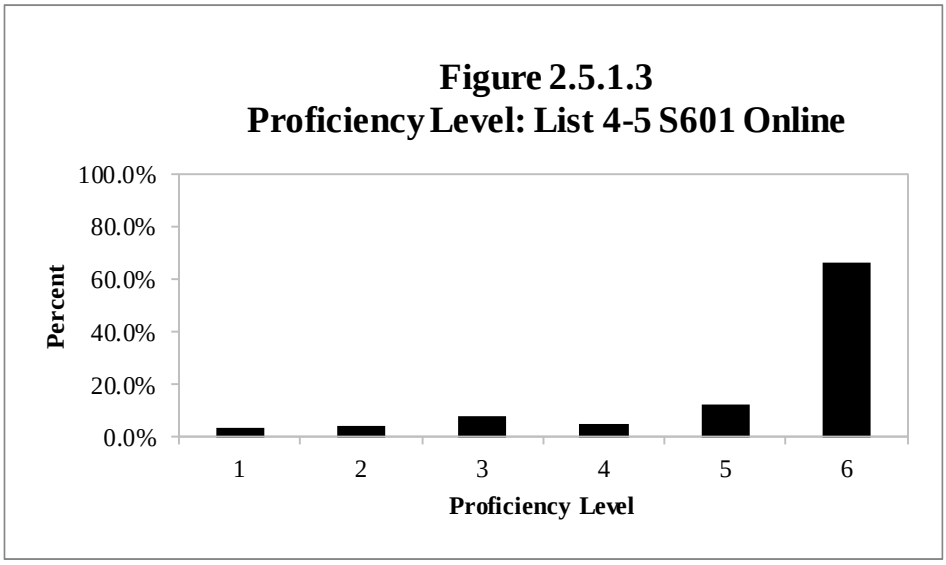
	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	32,690	15.73%	30,056	14.50%	62,746	15.12%
2	24,900	11.98%	23,316	11.25%	48,216	11.62%
3	41,251	19.85%	41,991	20.26%	83,242	20.06%
4	21,771	10.47%	22,717	10.96%	44,488	10.72%
5	37,948	18.26%	35,042	16.91%	72,990	17.59%
6	49,281	23.71%	54,096	26.11%	103,377	24.91%
Total	207,841	100.00%	207,218	100.00%	415,059	100.00%



2.5.1.3 Grade 4-5

Table 2.5.1.3
Proficiency Level Distribution: List 4-5 S601 Online

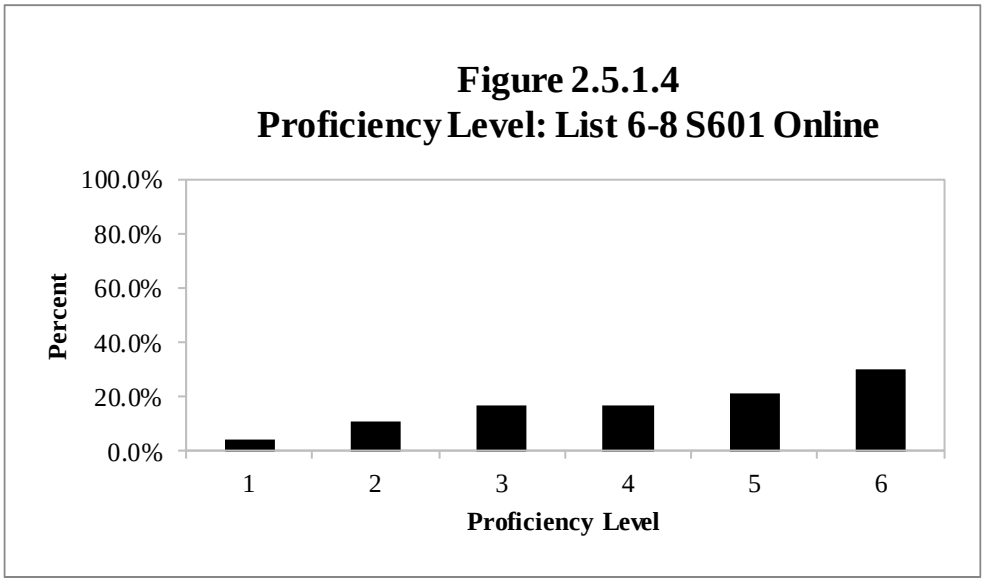
	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	4,606	2.31%	8,071	4.87%	12,677	3.47%
2	8,044	4.04%	8,729	5.26%	16,773	4.60%
3	15,851	7.96%	13,242	7.99%	29,093	7.97%
4	9,413	4.73%	9,857	5.94%	19,270	5.28%
5	22,518	11.32%	22,922	13.82%	45,440	12.46%
6	138,578	69.63%	102,999	62.11%	241,577	66.22%
Total	199,010	100.00%	165,820	100.00%	364,830	100.00%



2.5.1.4 Grade 6-8

Table 2.5.1.4
Proficiency Level Distribution: List 6-8 S601 Online

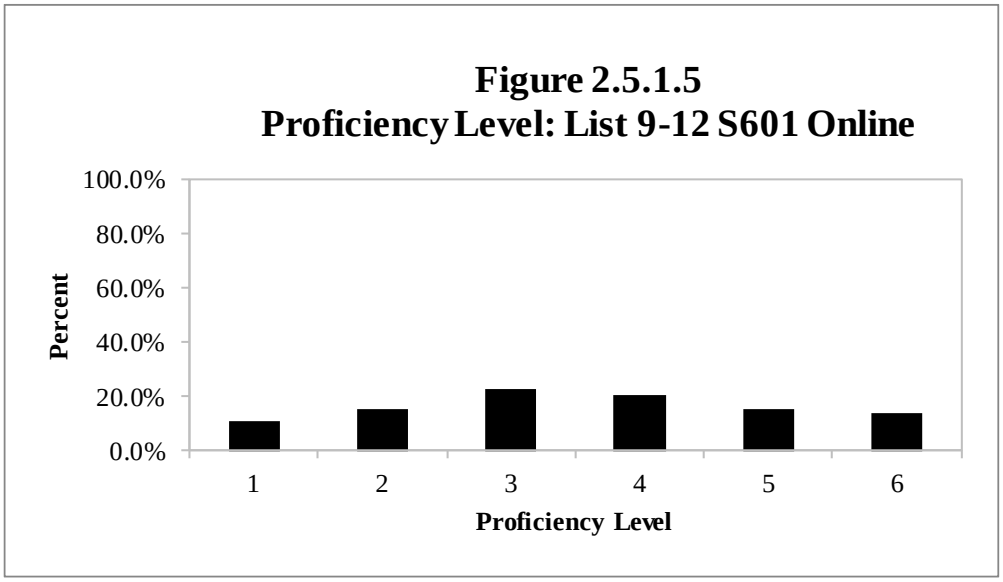
	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	3,563	2.54%	5,803	4.14%	8,536	6.39%	17,902	4.33%
2	13,721	9.79%	15,072	10.75%	15,618	11.70%	44,411	10.73%
3	22,759	16.24%	23,596	16.83%	22,427	16.80%	68,782	16.62%
4	23,019	16.43%	23,588	16.83%	23,053	17.27%	69,660	16.83%
5	33,957	24.23%	28,895	20.61%	26,377	19.76%	89,229	21.56%
6	43,101	30.76%	43,217	30.83%	37,488	28.08%	123,806	29.92%
Total	140,120	100.00%	140,171	100.00%	133,499	100.00%	413,790	100.00%



2.5.1.5 Grade 9-12

Table 2.5.1.5
Proficiency Level Distribution: List 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	13,531	8.75%	13,266	10.55%	12,281	13.51%	11,027	14.81%	50,105	11.24%
2	27,227	17.60%	18,670	14.84%	11,659	12.83%	10,277	13.80%	67,833	15.22%
3	35,305	22.83%	31,437	24.99%	20,125	22.14%	16,615	22.31%	103,482	23.21%
4	30,679	19.84%	25,180	20.02%	19,535	21.49%	16,631	22.33%	92,025	20.64%
5	24,533	15.86%	19,440	15.45%	15,621	17.18%	10,373	13.93%	69,967	15.69%
6	23,382	15.12%	17,795	14.15%	11,680	12.85%	9,546	12.82%	62,403	14.00%
Total	154,657	100.00%	125,788	100.00%	90,901	100.00%	74,469	100.00%	445,815	100.00%



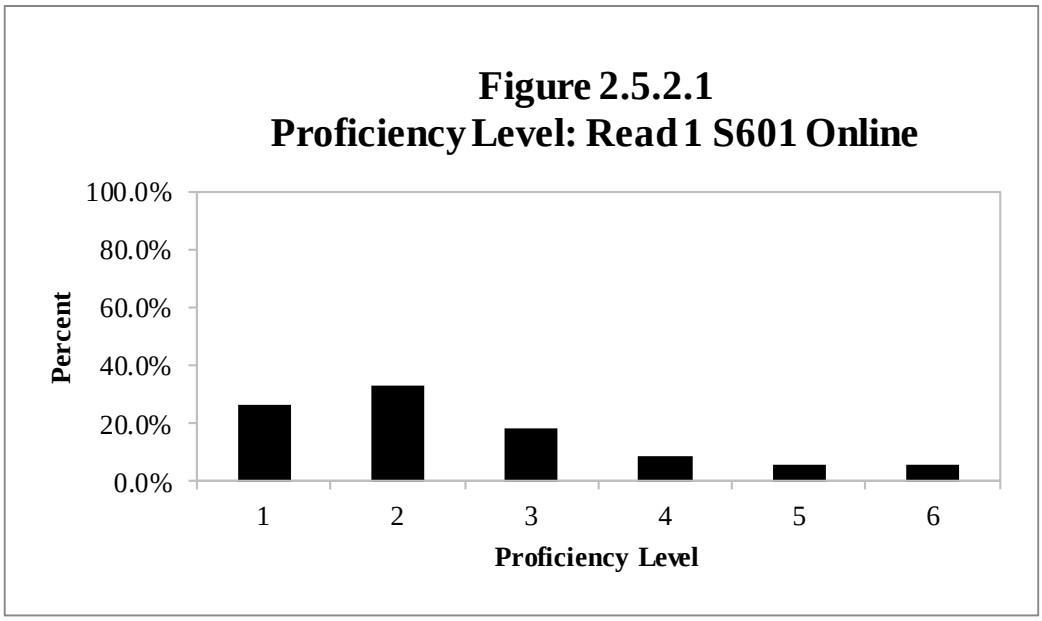
2.5.2 Reading

2.5.2.1 Grade 1

Table 2.5.2.1

Proficiency Level Distribution: Read 1 S601 Online

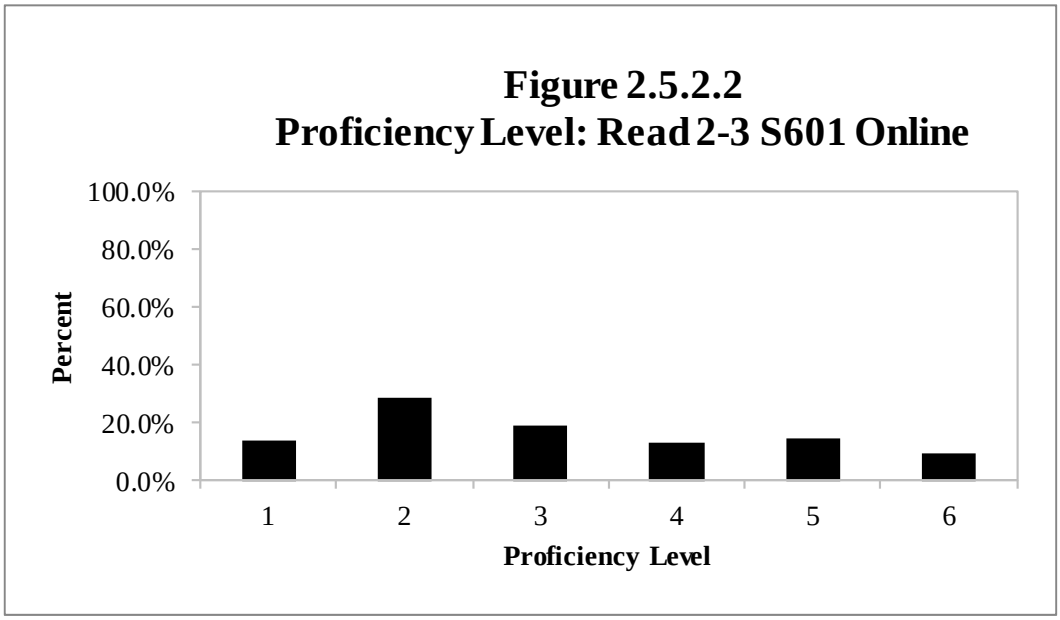
Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	58,284	26.63%	58,284	26.63%
2	73,543	33.60%	73,543	33.60%
3	40,204	18.37%	40,204	18.37%
4	19,911	9.10%	19,911	9.10%
5	13,386	6.12%	13,386	6.12%
6	13,568	6.20%	13,568	6.20%
Total	218,896	100.00%	218,896	100.00%



2.5.2.2 Grade 2-3

Table 2.5.2.2
Proficiency Level Distribution: Read 2-3 S601 Online

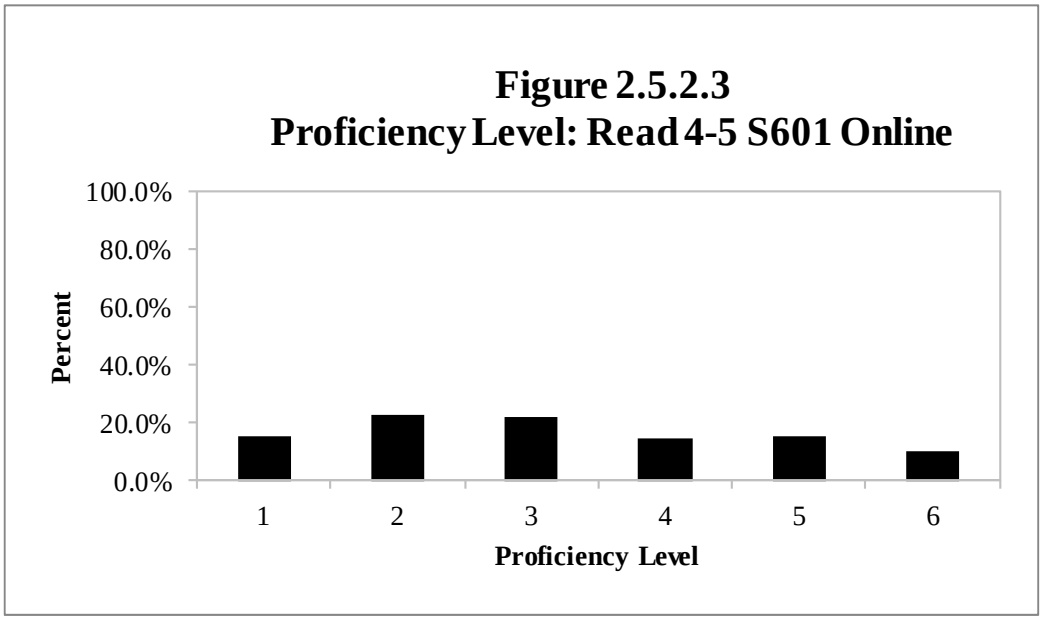
	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	19,919	9.34%	38,615	18.39%	58,534	13.83%
2	59,654	27.96%	60,976	29.04%	120,630	28.50%
3	50,092	23.48%	30,813	14.67%	80,905	19.11%
4	33,017	15.48%	24,323	11.58%	57,340	13.55%
5	34,105	15.99%	29,863	14.22%	63,968	15.11%
6	16,535	7.75%	25,415	12.10%	41,950	9.91%
Total	213,322	100.00%	210,005	100.00%	423,327	100.00%



2.5.2.3 Grade 4-5

Table 2.5.2.3
Proficiency Level Distribution: Read 4-5 S601 Online

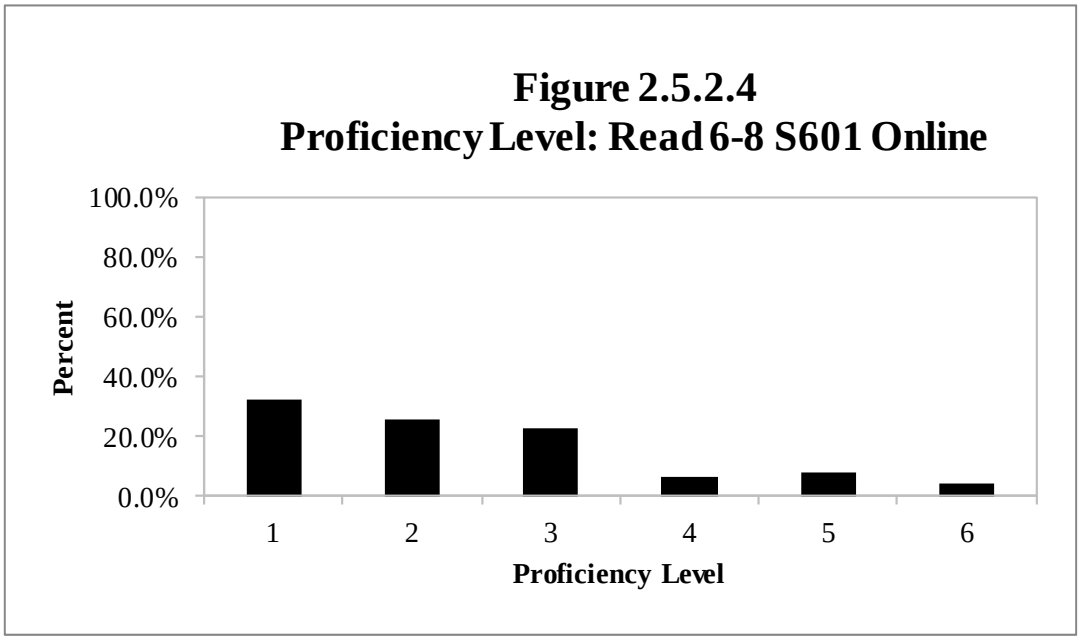
	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	25,190	12.70%	30,014	18.22%	55,204	15.20%
2	44,423	22.39%	37,810	22.95%	82,233	22.65%
3	39,960	20.14%	39,533	24.00%	79,493	21.89%
4	33,784	17.03%	18,647	11.32%	52,431	14.44%
5	32,267	16.27%	23,508	14.27%	55,775	15.36%
6	22,759	11.47%	15,208	9.23%	37,967	10.46%
Total	198,383	100.00%	164,720	100.00%	363,103	100.00%



2.5.2.4 Grade 6-8

Table 2.5.2.4
Proficiency Level Distribution: Read 6-8 S601 Online

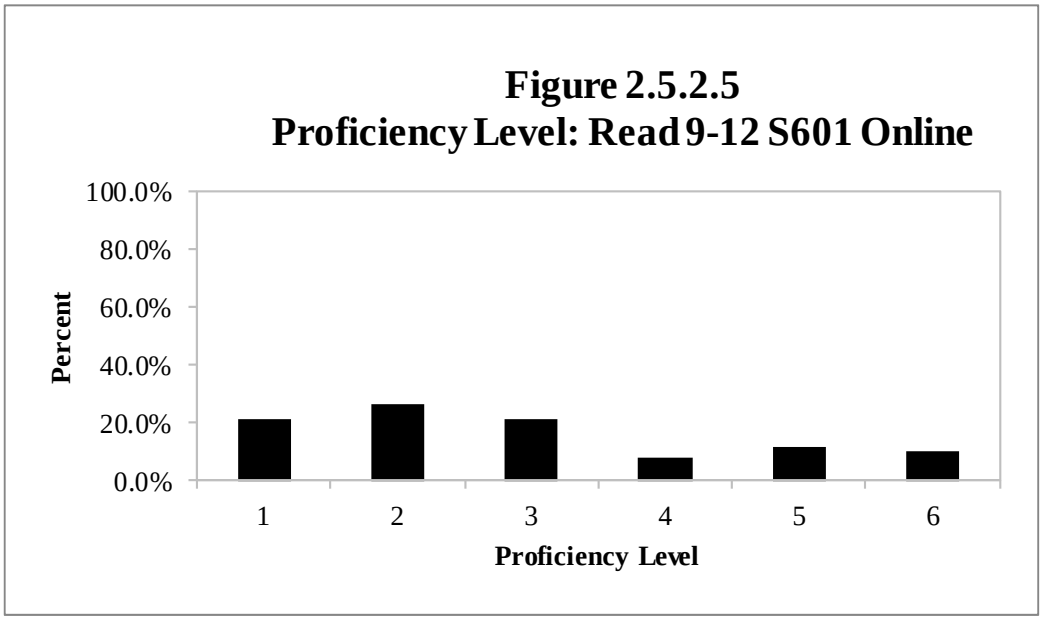
	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	42,595	30.12%	45,116	31.91%	46,464	34.64%	134,175	32.18%
2	39,111	27.66%	37,383	26.44%	31,758	23.68%	108,252	25.96%
3	35,403	25.03%	31,546	22.31%	27,654	20.62%	94,603	22.69%
4	9,894	7.00%	9,078	6.42%	10,027	7.47%	28,999	6.95%
5	10,536	7.45%	11,998	8.49%	10,610	7.91%	33,144	7.95%
6	3,878	2.74%	6,275	4.44%	7,628	5.69%	17,781	4.26%
Total	141,417	100.00%	141,396	100.00%	134,141	100.00%	416,954	100.00%



2.5.2.5 Grade 9-12

Table 2.5.2.5
Proficiency Level Distribution: Read 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	32,298	21.09%	26,278	21.19%	18,237	20.40%	17,055	23.21%	93,868	21.33%
2	41,625	27.18%	33,177	26.75%	22,592	25.27%	19,098	25.99%	116,492	26.47%
3	32,014	20.91%	25,454	20.52%	19,962	22.32%	16,953	23.07%	94,383	21.45%
4	12,074	7.88%	12,091	9.75%	7,512	8.40%	5,559	7.56%	37,236	8.46%
5	19,329	12.62%	14,895	12.01%	11,252	12.58%	8,031	10.93%	53,507	12.16%
6	15,796	10.32%	12,137	9.79%	9,862	11.03%	6,790	9.24%	44,585	10.13%
Total	153,136	100.00%	124,032	100.00%	89,417	100.00%	73,486	100.00%	440,071	100.00%



2.5.3 Writing

2.5.3.1 Grade 1

Table 2.5.3.1.1

Proficiency Level Distribution: Writ 1 A S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	101,900	49.55%	101,900	49.55%
2	72,447	35.23%	72,447	35.23%
3	31,282	15.21%	31,282	15.21%
4	24	0.01%	24	0.01%
5	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%
Total	205,653	100.00%	205,653	100.00%

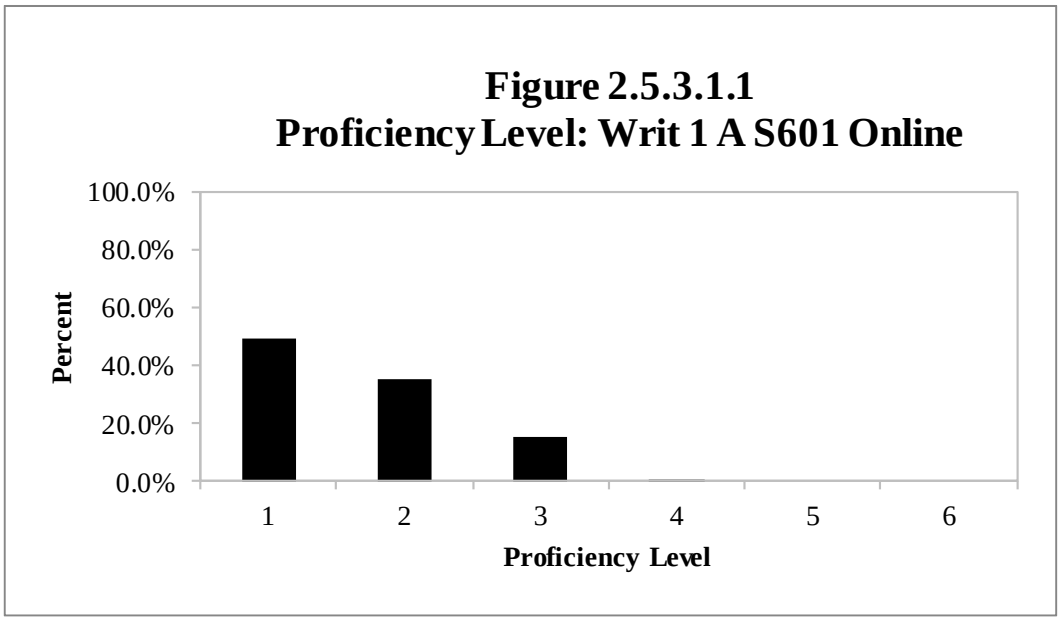


Table 2.5.3.1.2

Proficiency Level Distribution: Writ 1 B/C S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	604	2.72%	604	2.72%
2	4,620	20.81%	4,620	20.81%
3	14,557	65.58%	14,557	65.58%
4	2,403	10.83%	2,403	10.83%
5	8	0.04%	8	0.04%
6	5	0.02%	5	0.02%
Total	22,197	100.00%	22,197	100.00%

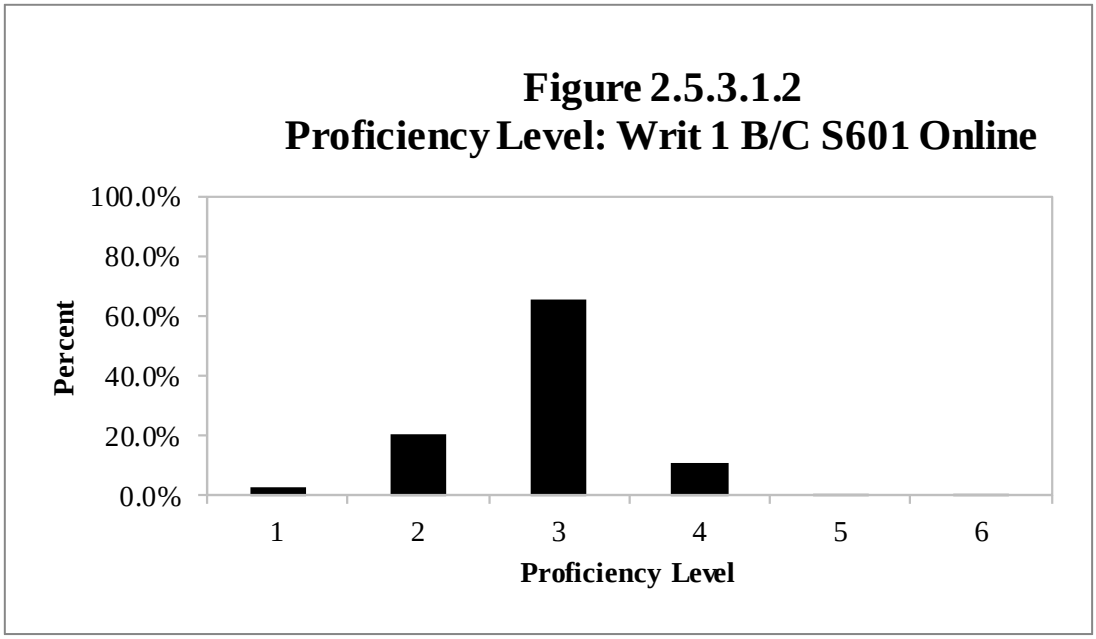
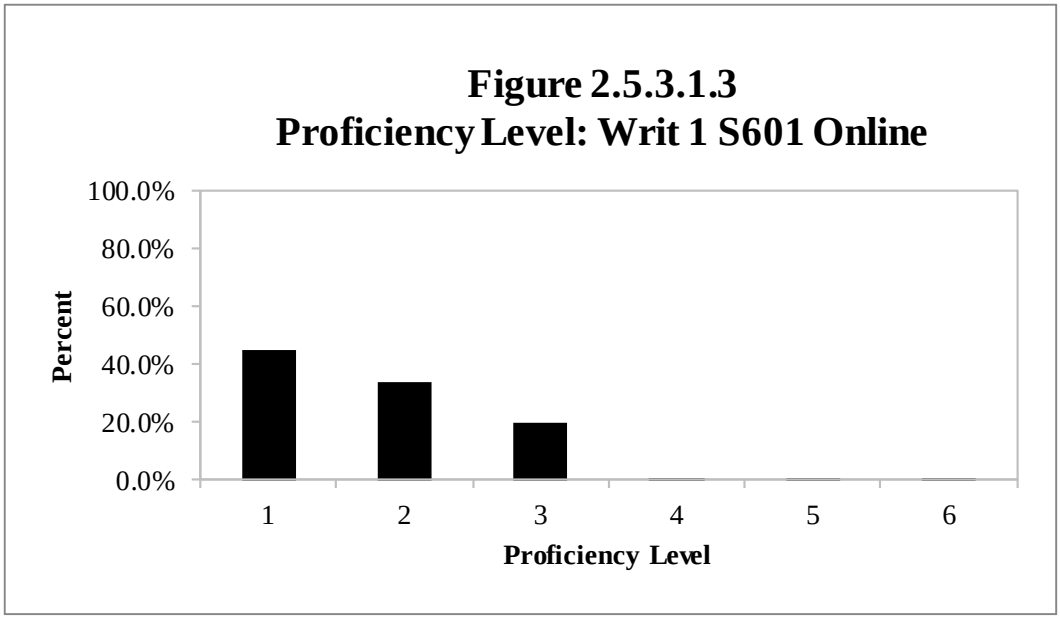


Table 2.5.3.1.3

Proficiency Level Distribution: Writ 1 S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	102,504	44.99%	102,504	44.99%
2	77,067	33.82%	77,067	33.82%
3	45,839	20.12%	45,839	20.12%
4	2,427	1.07%	2,427	1.07%
5	8	0.00%	8	0.00%
6	5	0.00%	5	0.00%
Total	227,850	100.00%	227,850	100.00%



2.5.3.2 Grade 2-3

Table 2.5.3.2.1
Proficiency Level Distribution: Writ 2-3 A S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	31,196	41.60%	24,534	39.42%	55,730	40.61%
2	32,173	42.90%	22,952	36.88%	55,125	40.17%
3	11,323	15.10%	14,598	23.46%	25,921	18.89%
4	295	0.39%	152	0.24%	447	0.33%
5	0	0.00%	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%	0	0.00%
Total	74,987	100.00%	62,236	100.00%	137,223	100.00%

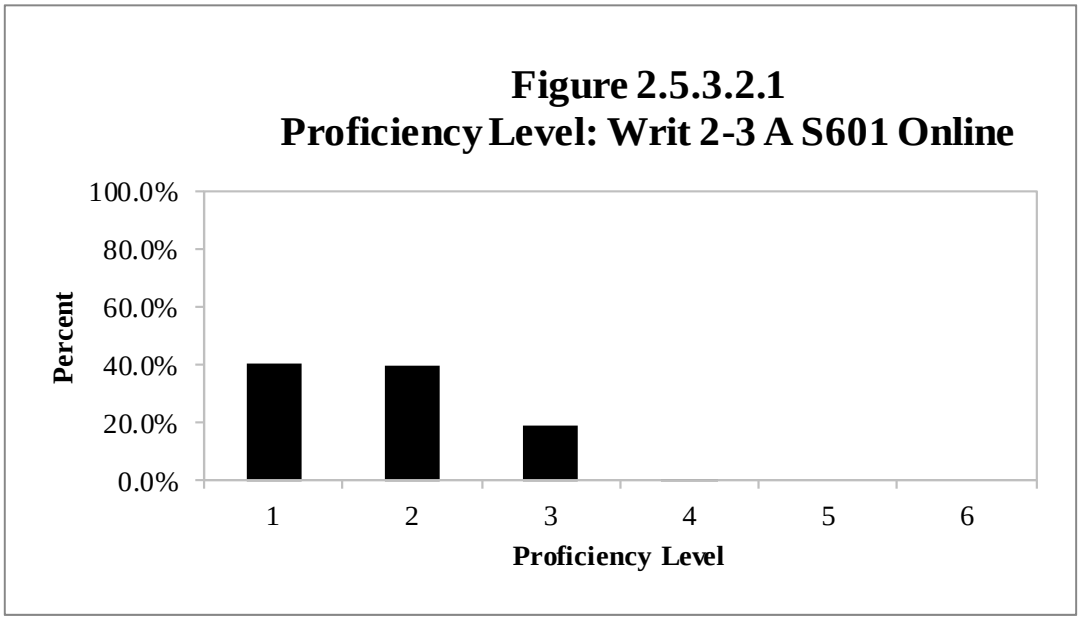


Table 2.5.3.2.2

Proficiency Level Distribution: Writ 2-3 B/C S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	10,680	7.14%	4,438	2.79%	15,118	4.90%
2	31,854	21.30%	13,748	8.64%	45,602	14.77%
3	98,304	65.74%	113,514	71.33%	211,818	68.62%
4	8,640	5.78%	26,985	16.96%	35,625	11.54%
5	56	0.04%	452	0.28%	508	0.16%
6	0	0.00%	11	0.01%	11	0.00%
Total	149,534	100.00%	159,148	100.00%	308,682	100.00%

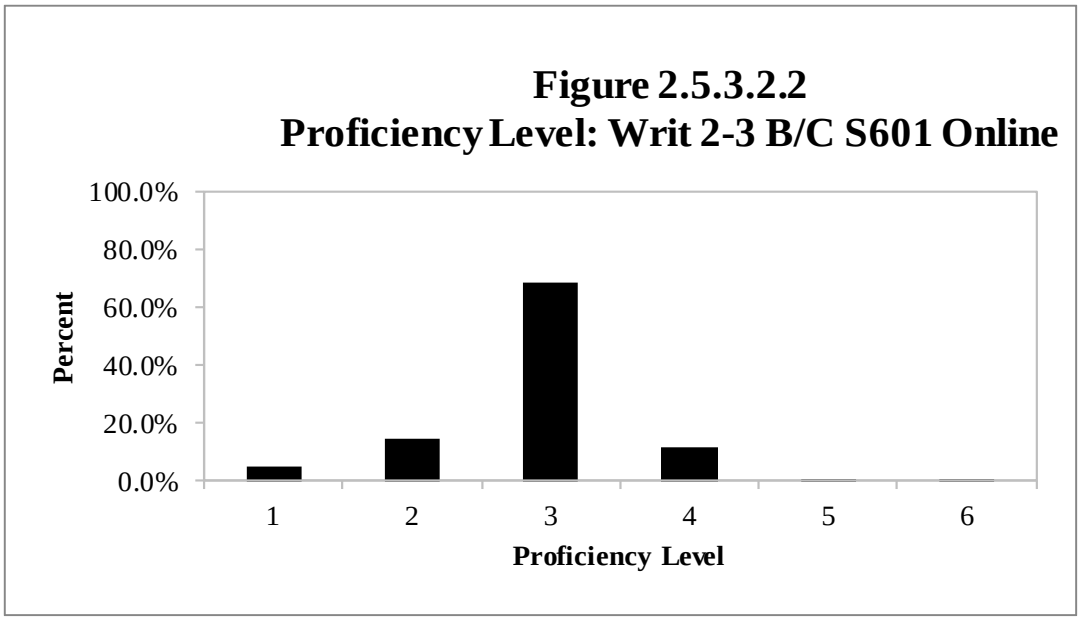
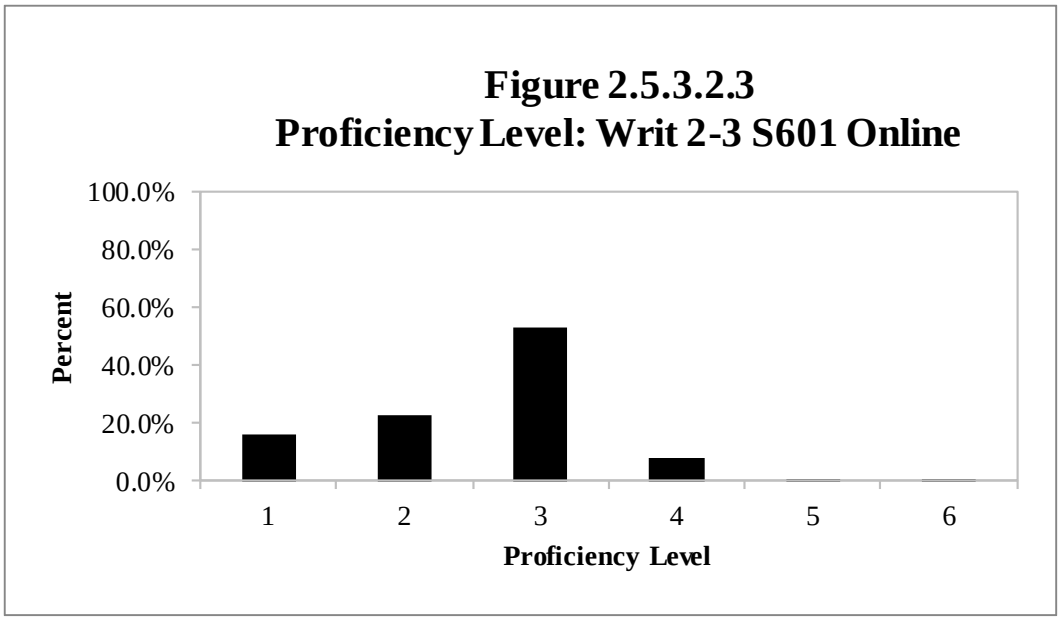


Table 2.5.3.2.3

Proficiency Level Distribution: Writ 2-3 S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	41,876	18.65%	28,972	13.09%	70,848	15.89%
2	64,027	28.52%	36,700	16.58%	100,727	22.59%
3	109,627	48.83%	128,112	57.87%	237,739	53.32%
4	8,935	3.98%	27,137	12.26%	36,072	8.09%
5	56	0.02%	452	0.20%	508	0.11%
6	0	0.00%	11	0.00%	11	0.00%
Total	224,521	100.00%	221,384	100.00%	445,905	100.00%



2.5.3.3 Grade 4-5

Table 2.5.3.3.1
Proficiency Level Distribution: Writ 4-5 A S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	17,684	40.90%	17,074	39.52%	34,758	40.21%
2	13,385	30.96%	9,369	21.68%	22,754	26.32%
3	11,868	27.45%	15,947	36.91%	27,815	32.18%
4	299	0.69%	816	1.89%	1,115	1.29%
5	4	0.01%	1	0.00%	5	0.01%
6	0	0.00%	0	0.00%	0	0.00%
Total	43,240	100.00%	43,207	100.00%	86,447	100.00%

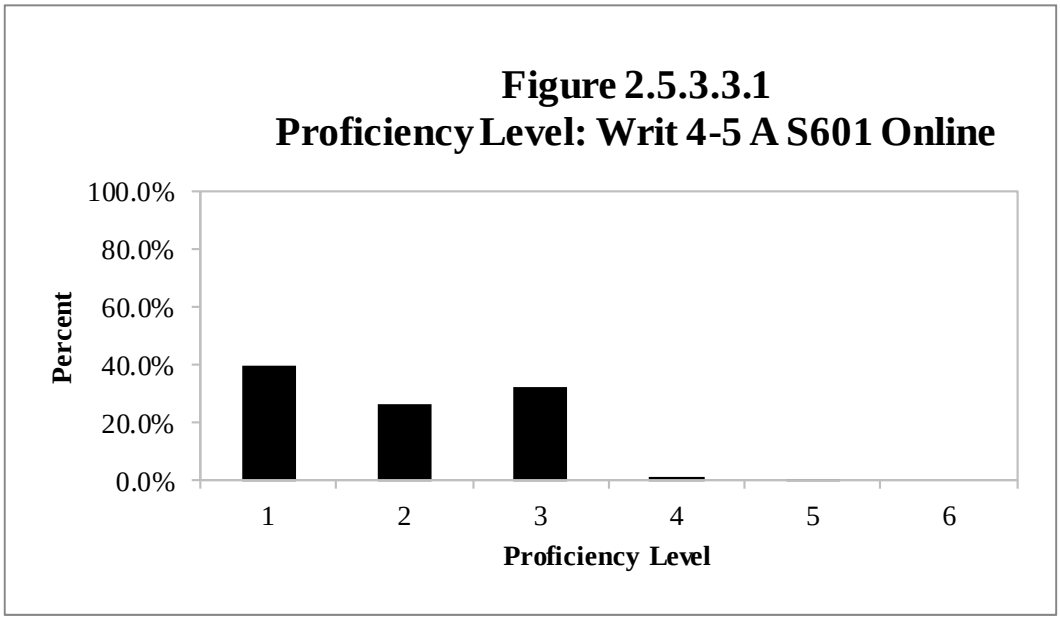


Table 2.5.3.3.2

Proficiency Level Distribution: Writ 4-5 B/C S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	2,591	1.62%	619	0.49%	3,210	1.12%
2	2,303	1.44%	1,584	1.26%	3,887	1.36%
3	85,793	53.59%	80,293	63.64%	166,086	58.02%
4	63,507	39.67%	40,999	32.50%	104,506	36.51%
5	5,523	3.45%	1,983	1.57%	7,506	2.62%
6	389	0.24%	682	0.54%	1,071	0.37%
Total	160,106	100.00%	126,160	100.00%	286,266	100.00%

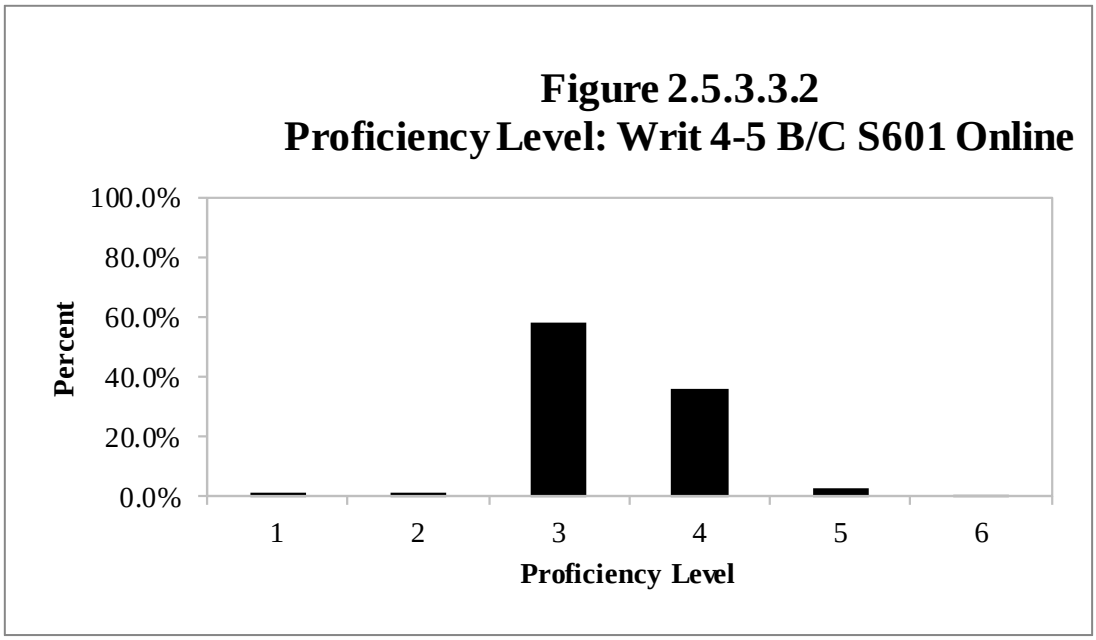
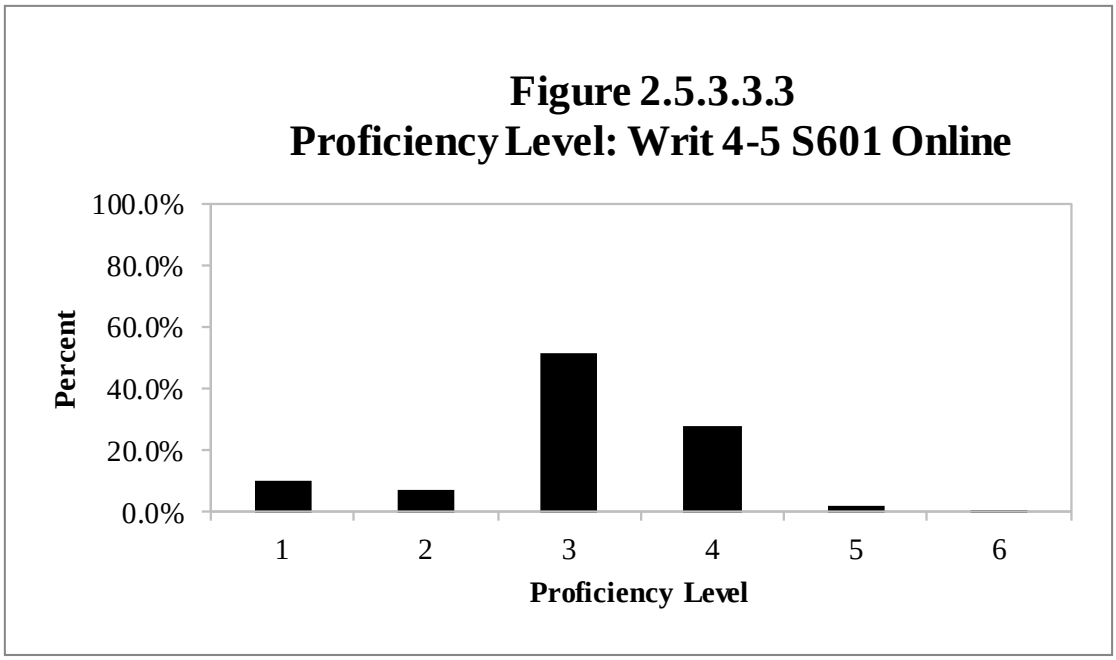


Table 2.5.3.3.3

Proficiency Level Distribution: Writ 4-5 S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	20,275	9.97%	17,693	10.45%	37,968	10.19%
2	15,688	7.71%	10,953	6.47%	26,641	7.15%
3	97,661	48.03%	96,240	56.82%	193,901	52.02%
4	63,806	31.38%	41,815	24.69%	105,621	28.34%
5	5,527	2.72%	1,984	1.17%	7,511	2.02%
6	389	0.19%	682	0.40%	1,071	0.29%
Total	203,346	100.00%	169,367	100.00%	372,713	100.00%



2.5.3.4 Grade 6-8

Table 2.5.3.4.1
Proficiency Level Distribution: Writ 6-8 A S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	16,923	34.40%	22,112	36.49%	20,129	31.53%	59,164	34.07%
2	20,005	40.66%	16,624	27.43%	15,511	24.30%	52,140	30.03%
3	11,629	23.64%	21,419	35.34%	27,291	42.75%	60,339	34.75%
4	636	1.29%	443	0.73%	911	1.43%	1,990	1.15%
5	2	0.00%	5	0.01%	1	0.00%	8	0.00%
6	0	0.00%	1	0.00%	1	0.00%	2	0.00%
Total	49,195	100.00%	60,604	100.00%	63,844	100.00%	173,643	100.00%

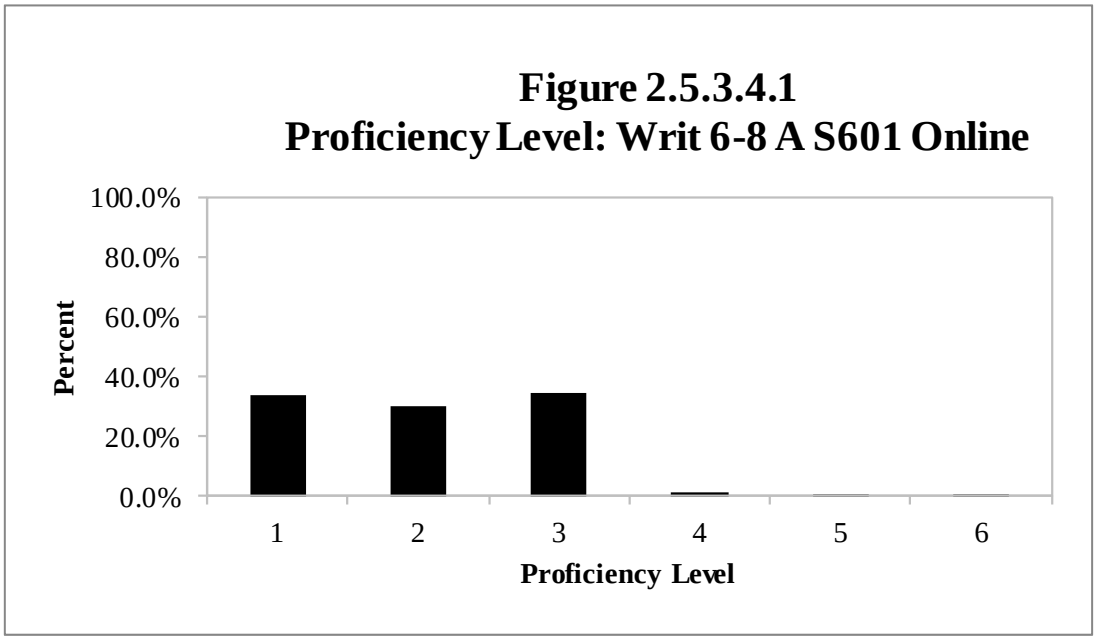


Table 2.5.3.4.2

Proficiency Level Distribution: Writ 6-8 B/C S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	1,964	2.04%	1,199	1.41%	941	1.28%	4,104	1.61%
2	16,561	17.17%	6,588	7.77%	8,615	11.74%	31,764	12.47%
3	63,718	66.07%	68,549	80.80%	51,514	70.23%	183,781	72.17%
4	14,186	14.71%	8,439	9.95%	12,259	16.71%	34,884	13.70%
5	14	0.01%	64	0.08%	24	0.03%	102	0.04%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	96,443	100.00%	84,839	100.00%	73,353	100.00%	254,635	100.00%

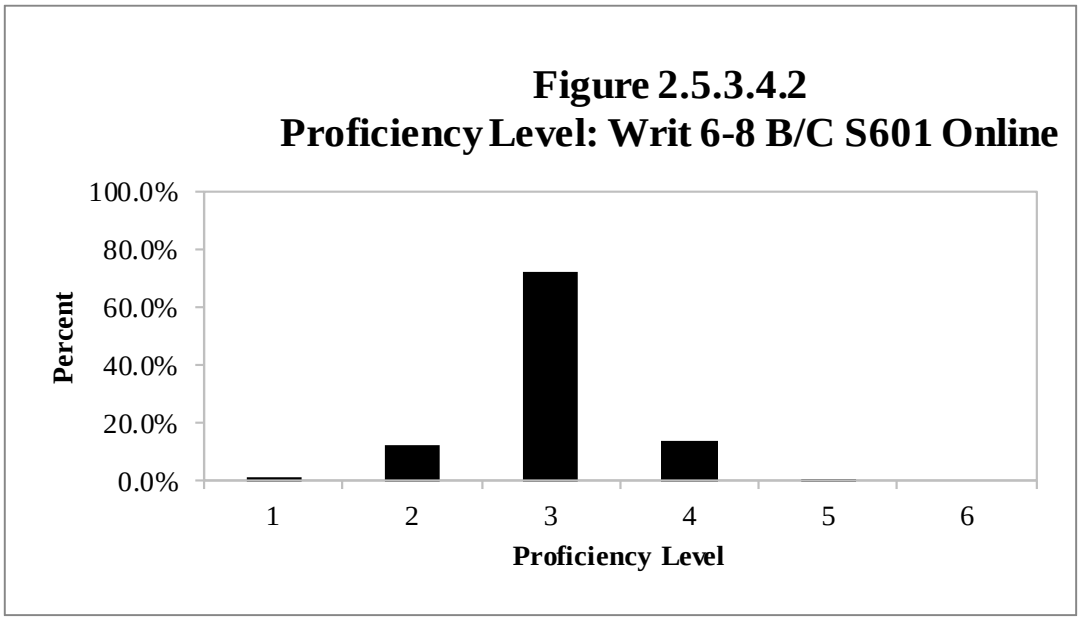
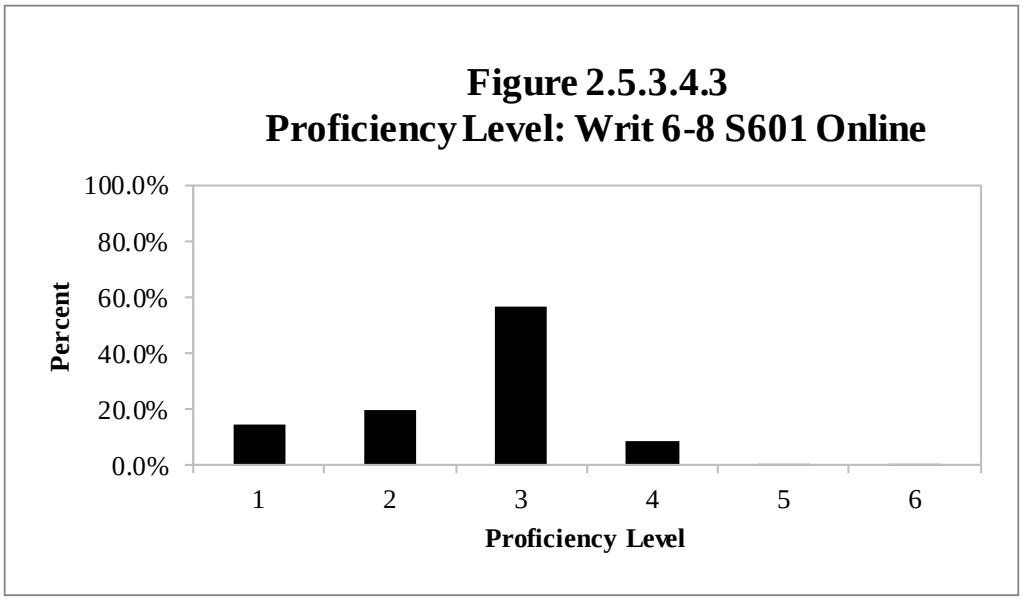


Table 2.5.3.4.3

Proficiency Level Distribution: Writ 6-8 S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	18,887	12.97%	23,311	16.03%	21,070	15.36%	63,268	14.77%
2	36,566	25.11%	23,212	15.96%	24,126	17.58%	83,904	19.59%
3	75,347	51.74%	89,968	61.86%	78,805	57.44%	244,120	57.00%
4	14,822	10.18%	8,882	6.11%	13,170	9.60%	36,874	8.61%
5	16	0.01%	69	0.05%	25	0.02%	110	0.03%
6	0	0.00%	1	0.00%	1	0.00%	2	0.00%
Total	145,638	100.00%	145,443	100.00%	137,197	100.00%	428,278	100.00%



2.5.3.5 Grade 9-12

Table 2.5.3.5.1
Proficiency Level Distribution: Writ 9-12 A S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	18,526	27.77%	12,817	25.91%	8,381	26.11%	8,953	36.94%	48,677	28.22%
2	17,844	26.75%	18,277	36.95%	14,482	45.12%	7,752	31.98%	58,355	33.83%
3	26,241	39.33%	17,280	34.93%	8,265	25.75%	6,699	27.64%	58,485	33.90%
4	4,011	6.01%	1,078	2.18%	953	2.97%	831	3.43%	6,873	3.98%
5	90	0.13%	13	0.03%	13	0.04%	2	0.01%	118	0.07%
6	2	0.00%	1	0.00%	0	0.00%	0	0.00%	3	0.00%
Total	66,714	100.00%	49,466	100.00%	32,094	100.00%	24,237	100.00%	172,511	100.00%

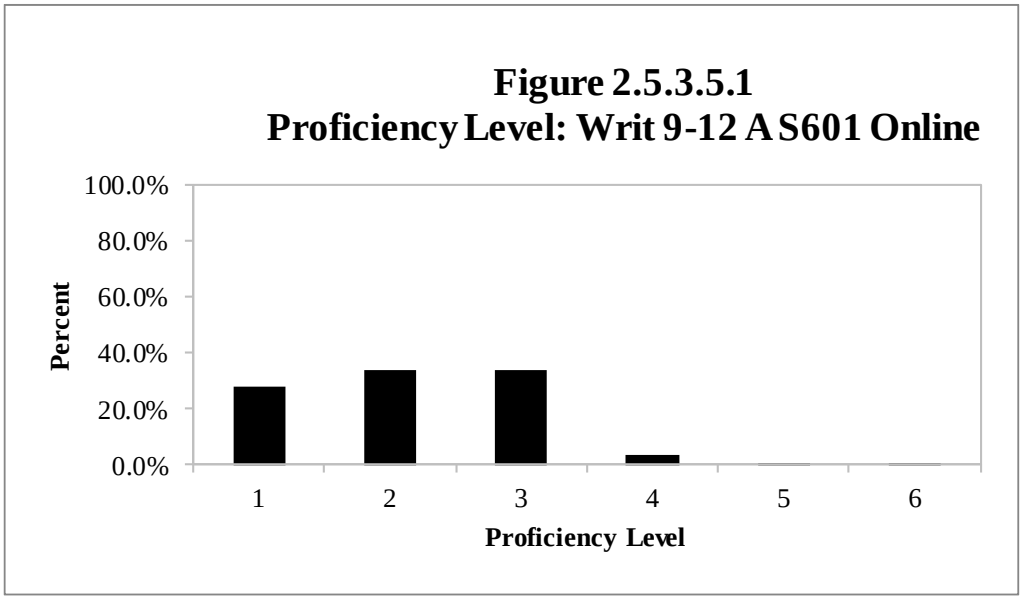


Table 2.5.3.5.2

Proficiency Level Distribution: Writ 9-12 B/C S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	1,512	1.64%	2,159	2.73%	3,386	5.60%	6,457	12.49%	13,514	4.76%
2	11,229	12.16%	8,257	10.42%	10,853	17.94%	6,851	13.26%	37,190	13.11%
3	53,031	57.43%	57,632	72.75%	35,732	59.07%	29,942	57.93%	176,337	62.15%
4	26,128	28.30%	10,751	13.57%	10,395	17.18%	8,361	16.18%	55,635	19.61%
5	412	0.45%	423	0.53%	125	0.21%	71	0.14%	1,031	0.36%
6	22	0.02%	1	0.00%	3	0.00%	2	0.00%	28	0.01%
Total	92,334	100.00%	79,223	100.00%	60,494	100.00%	51,684	100.00%	283,735	100.00%

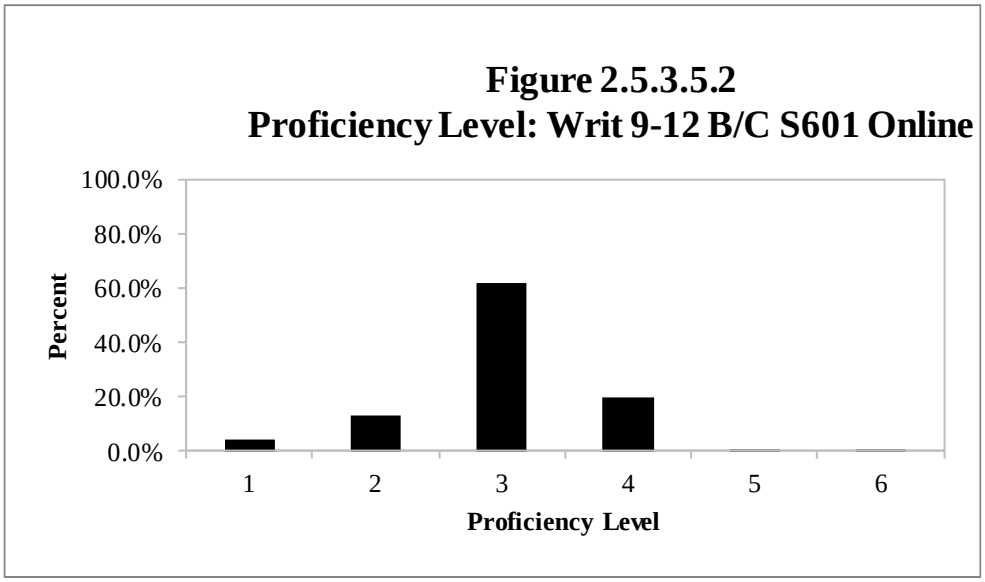
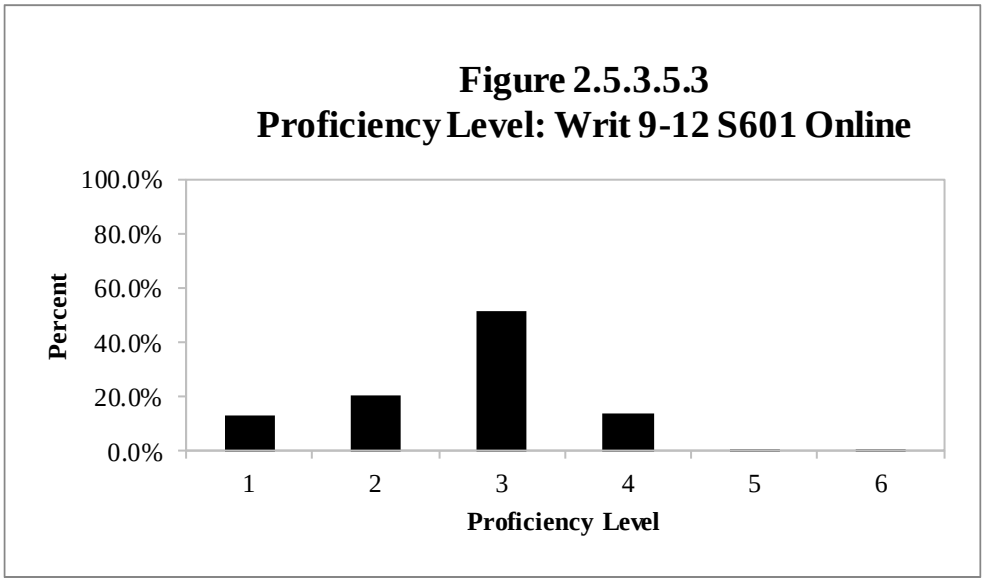


Table 2.5.3.5.3

Proficiency Level Distribution: Writ 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	20,038	12.60%	14,976	11.64%	11,767	12.71%	15,410	20.30%	62,191	13.63%
2	29,073	18.28%	26,534	20.62%	25,335	27.36%	14,603	19.23%	95,545	20.94%
3	79,272	49.84%	74,912	58.21%	43,997	47.52%	36,641	48.26%	234,822	51.47%
4	30,139	18.95%	11,829	9.19%	11,348	12.26%	9,192	12.11%	62,508	13.70%
5	502	0.32%	436	0.34%	138	0.15%	73	0.10%	1,149	0.25%
6	24	0.02%	2	0.00%	3	0.00%	2	0.00%	31	0.01%
Total	159,048	100.00%	128,689	100.00%	92,588	100.00%	75,921	100.00%	456,246	100.00%



2.5.4 Speaking

2.5.4.1 Grade 1

Table 2.5.4.1.1

Proficiency Level Distribution: Spek 1 Pre-A S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	13,774	100.00%	13,774	100.00%
Total	13,774	100.00%	13,774	100.00%

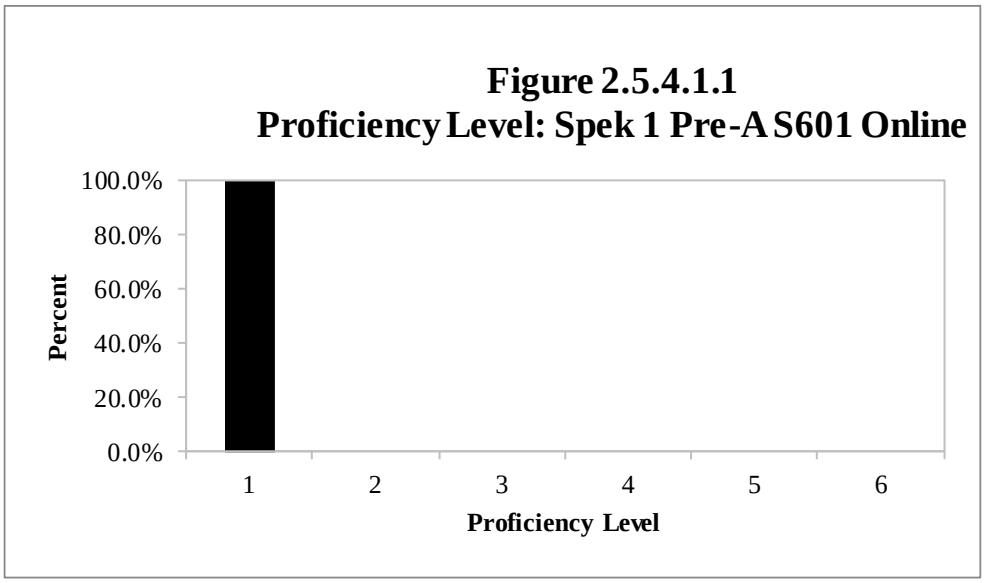


Table 2.5.4.1.2

Proficiency Level Distribution: Spek 1 A S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	27,648	27.16%	27,648	27.16%
2	50,984	50.09%	50,984	50.09%
3	18,520	18.20%	18,520	18.20%
4	4,577	4.50%	4,577	4.50%
5	50	0.05%	50	0.05%
6	0	0.00%	0	0.00%
Total	101,779	100.00%	101,779	100.00%

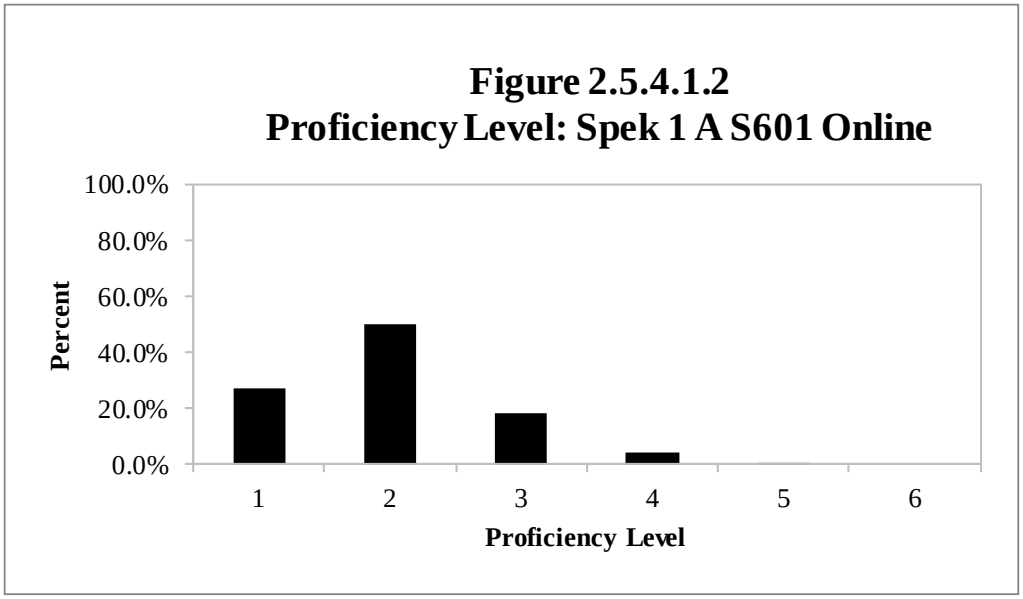


Table 2.5.4.1.3

Proficiency Level Distribution: Spek 1 B/C S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	6,841	7.20%	6,841	7.20%
2	31,008	32.64%	31,008	32.64%
3	39,657	41.74%	39,657	41.74%
4	16,636	17.51%	16,636	17.51%
5	781	0.82%	781	0.82%
6	81	0.09%	81	0.09%
Total	95,004	100.00%	95,004	100.00%

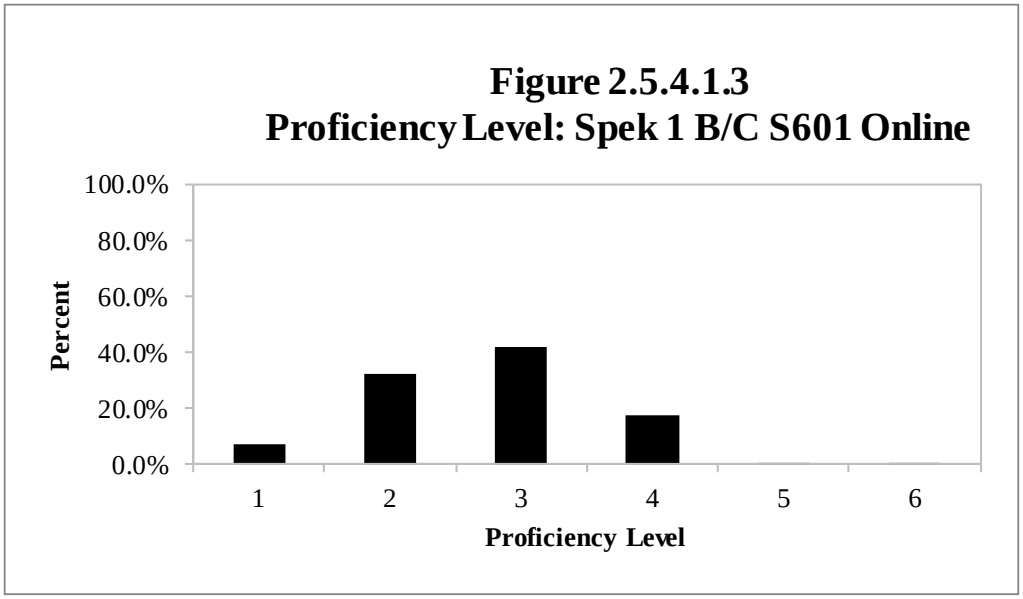
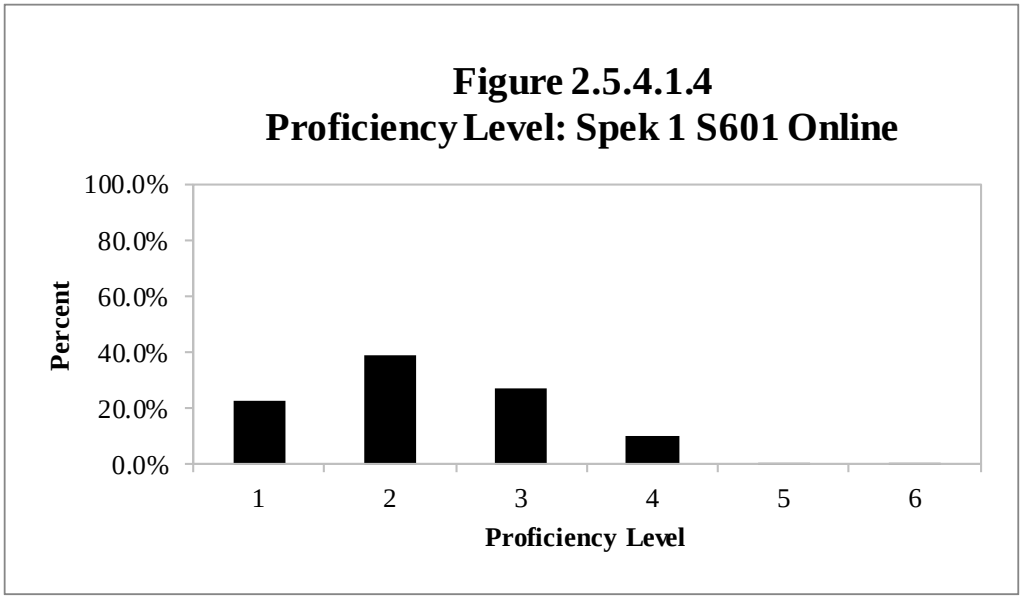


Table 2.5.4.1.4

Proficiency Level Distribution: Spek 1 S601 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	48,263	22.92%	48,263	22.92%
2	81,992	38.94%	81,992	38.94%
3	58,177	27.63%	58,177	27.63%
4	21,213	10.07%	21,213	10.07%
5	831	0.39%	831	0.39%
6	81	0.04%	81	0.04%
Total	210,557	100.00%	210,557	100.00%



2.5.4.2 Grade 2-3

Table 2.5.4.2.1
Proficiency Level Distribution: Spek 2-3 Pre-A S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	6,222	100.00%	14,627	100.00%	20,849	100.00%
Total	6,222	100.00%	14,627	100.00%	20,849	100.00%

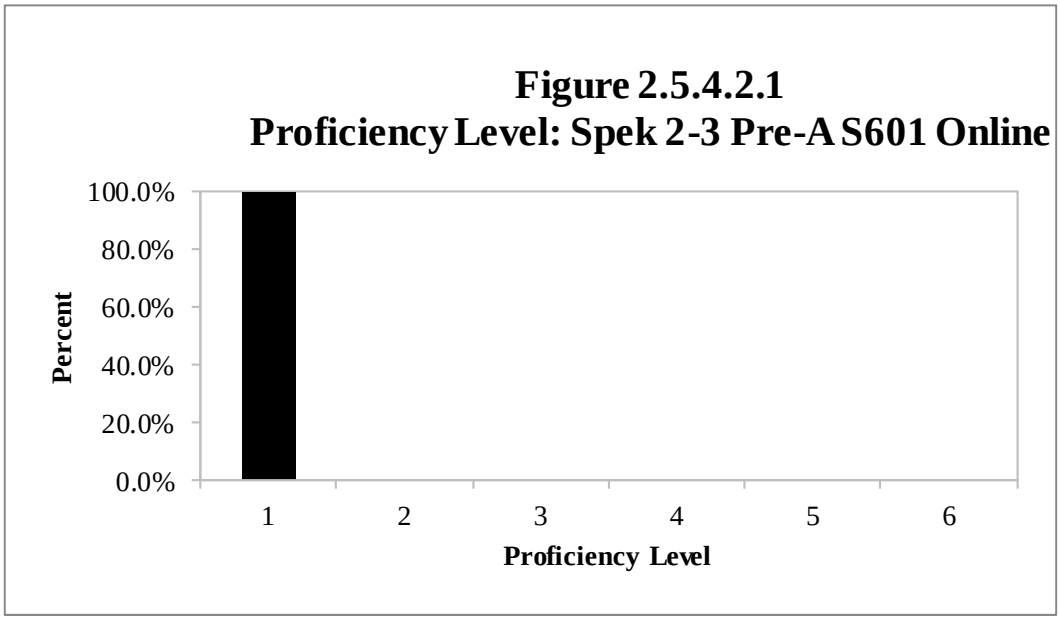


Table 2.5.4.2.2

Proficiency Level Distribution: Spek 2-3 A S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	20,513	33.50%	17,258	28.37%	37,771	30.94%
2	26,398	43.11%	18,645	30.65%	45,043	36.90%
3	11,963	19.54%	18,962	31.17%	30,925	25.33%
4	2,230	3.64%	5,900	9.70%	8,130	6.66%
5	124	0.20%	75	0.12%	199	0.16%
6	0	0.00%	0	0.00%	0	0.00%
Total	61,228	100.00%	60,840	100.00%	122,068	100.00%

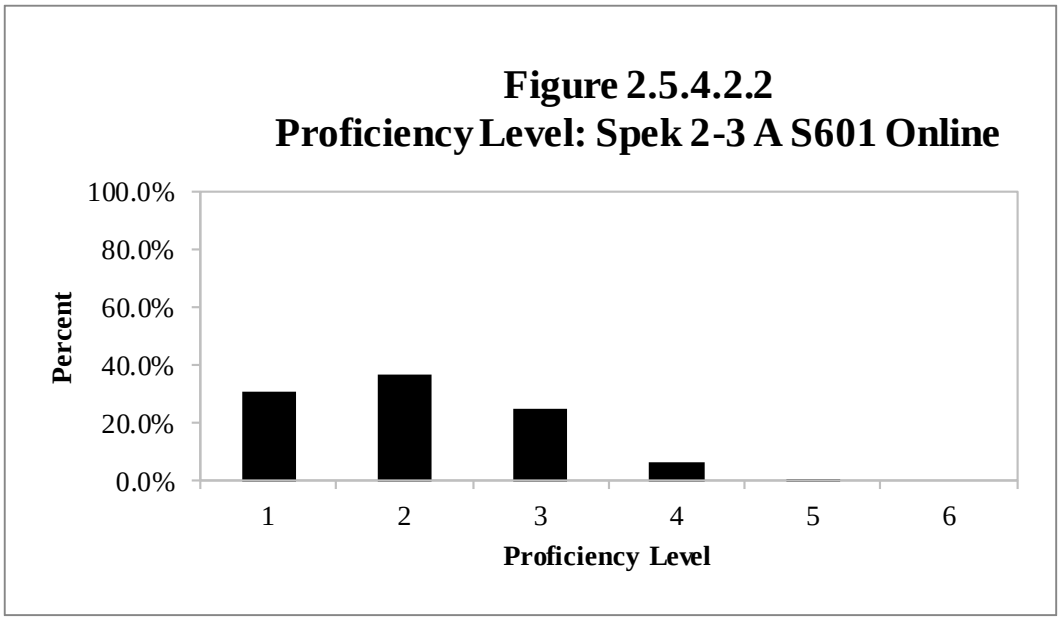


Table 2.5.4.2.3

Proficiency Level Distribution: Spek 2-3 B/C S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	11,194	8.02%	9,327	7.14%	20,521	7.60%
2	45,660	32.73%	30,855	23.63%	76,515	28.33%
3	55,885	40.05%	58,357	44.70%	114,242	42.30%
4	25,436	18.23%	30,575	23.42%	56,011	20.74%
5	1,224	0.88%	1,009	0.77%	2,233	0.83%
6	125	0.09%	427	0.33%	552	0.20%
Total	139,524	100.00%	130,550	100.00%	270,074	100.00%

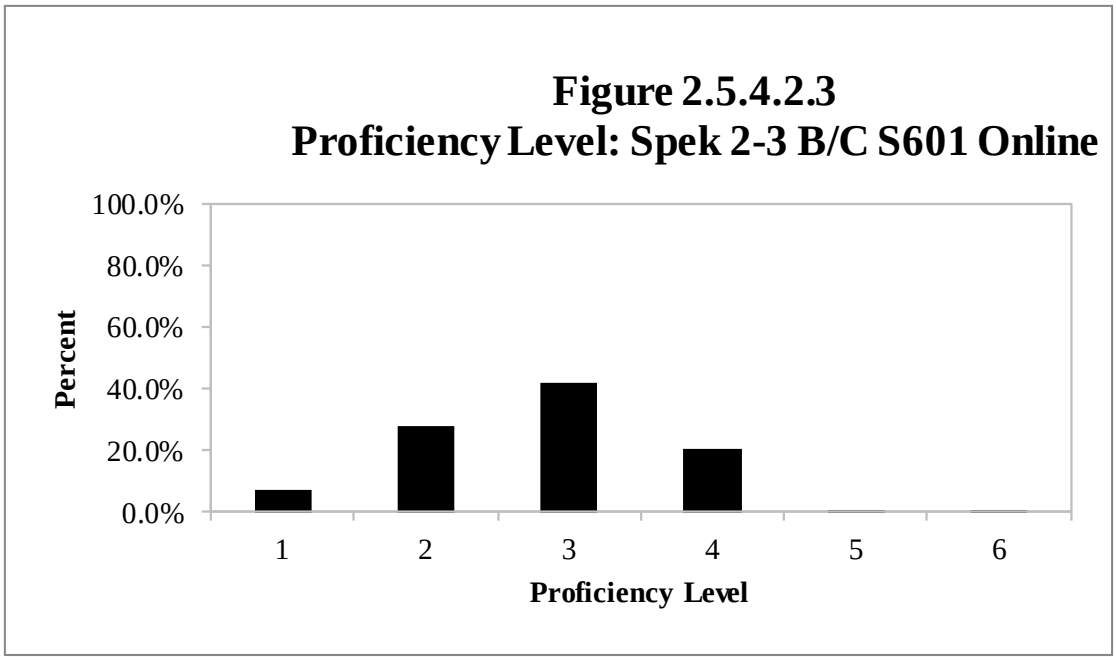
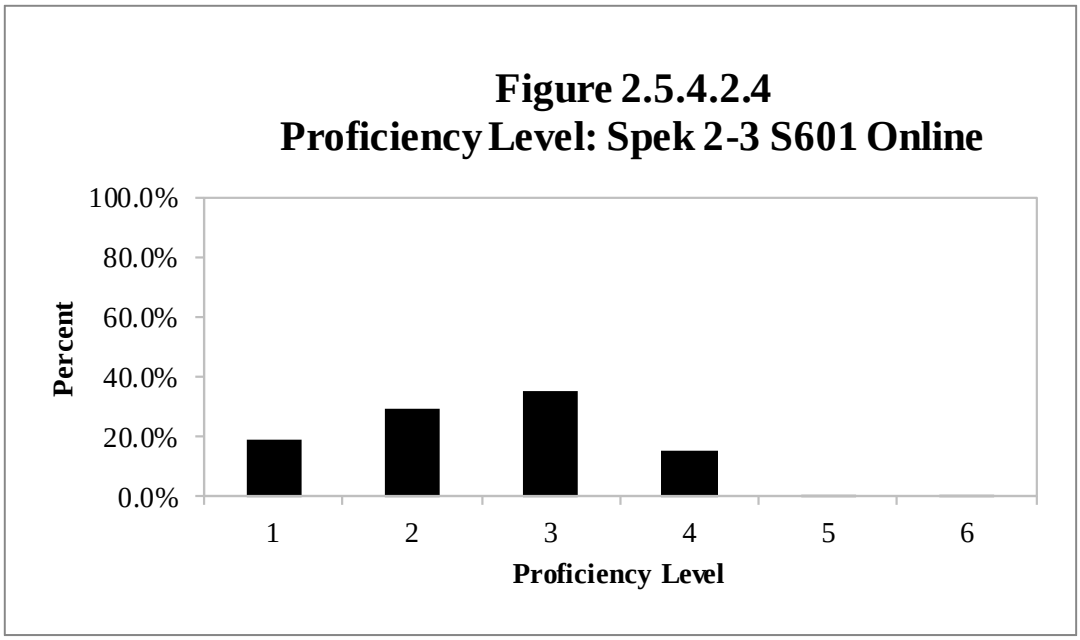


Table 2.5.4.2.4

Proficiency Level Distribution: Spek 2-3 S601 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	37,929	18.33%	41,212	20.00%	79,141	19.16%
2	72,058	34.82%	49,500	24.03%	121,558	29.43%
3	67,848	32.78%	77,319	37.53%	145,167	35.15%
4	27,666	13.37%	36,475	17.70%	64,141	15.53%
5	1,348	0.65%	1,084	0.53%	2,432	0.59%
6	125	0.06%	427	0.21%	552	0.13%
Total	206,974	100.00%	206,017	100.00%	412,991	100.00%



2.5.4.3 Grade 4-5

Table 2.5.4.3.1
Proficiency Level Distribution: Spek 4-5 Pre-A S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	2,426	100.00%	6,115	100.00%	8,541	100.00%
Total	2,426	100.00%	6,115	100.00%	8,541	100.00%

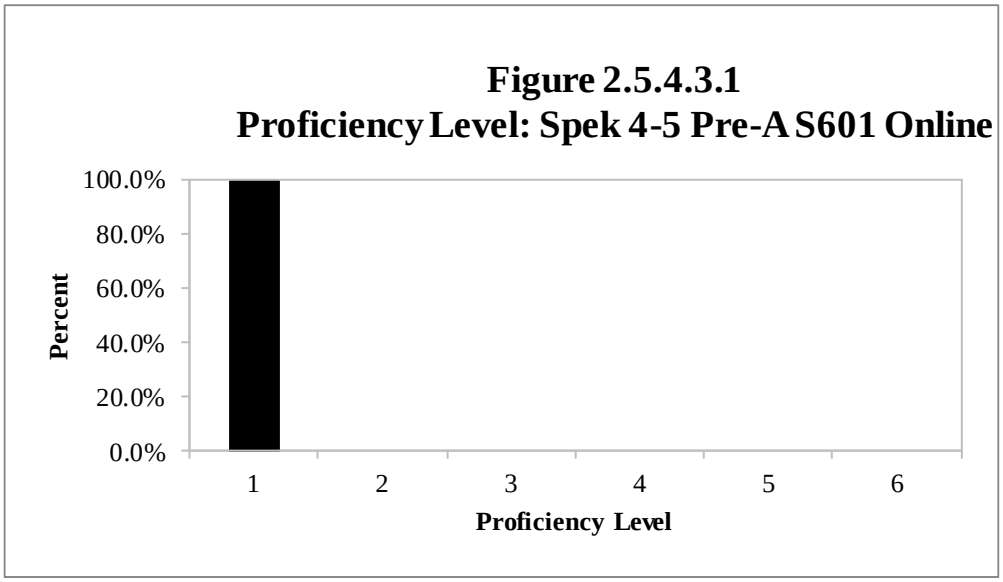


Table 2.5.4.3.2

Proficiency Level Distribution: Spek 4-5 A S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	13,175	40.93%	13,100	51.27%	26,275	45.50%
2	8,734	27.13%	7,527	29.46%	16,261	28.16%
3	7,611	23.64%	3,685	14.42%	11,296	19.56%
4	2,537	7.88%	1,083	4.24%	3,620	6.27%
5	135	0.42%	154	0.60%	289	0.50%
6	0	0.00%	0	0.00%	0	0.00%
Total	32,192	100.00%	25,549	100.00%	57,741	100.00%

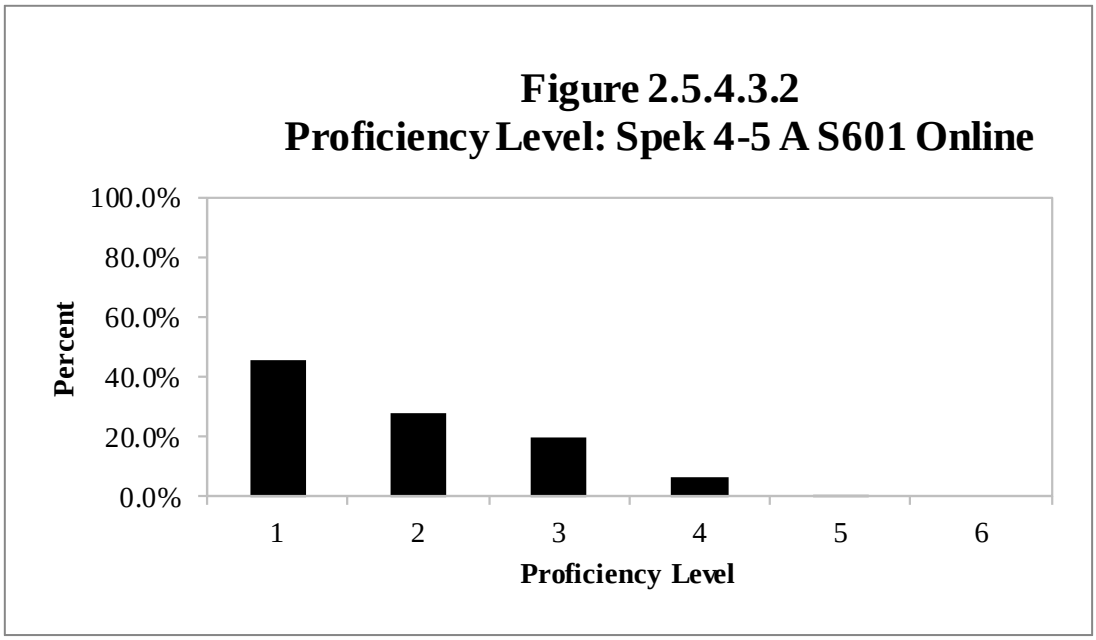


Table 2.5.4.3.3

Proficiency Level Distribution: Spek 4-5 B/C S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	3,797	2.33%	6,434	4.86%	10,231	3.46%
2	26,749	16.40%	25,516	19.28%	52,265	17.69%
3	60,476	37.07%	52,689	39.81%	113,165	38.30%
4	62,082	38.05%	42,649	32.22%	104,731	35.44%
5	9,075	5.56%	4,814	3.64%	13,889	4.70%
6	963	0.59%	258	0.19%	1,221	0.41%
Total	163,142	100.00%	132,360	100.00%	295,502	100.00%

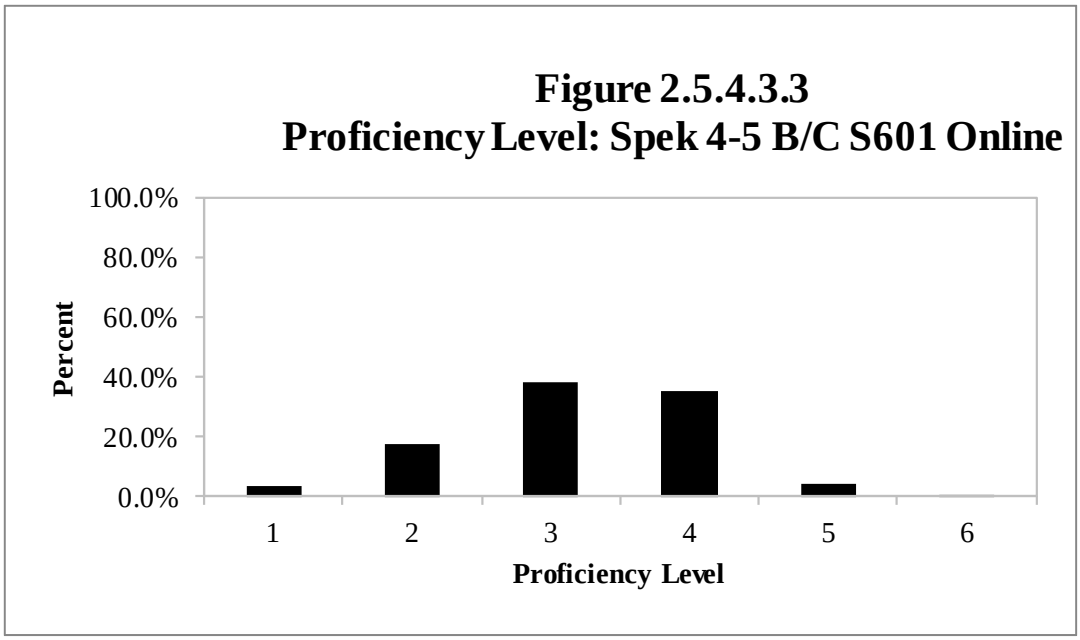
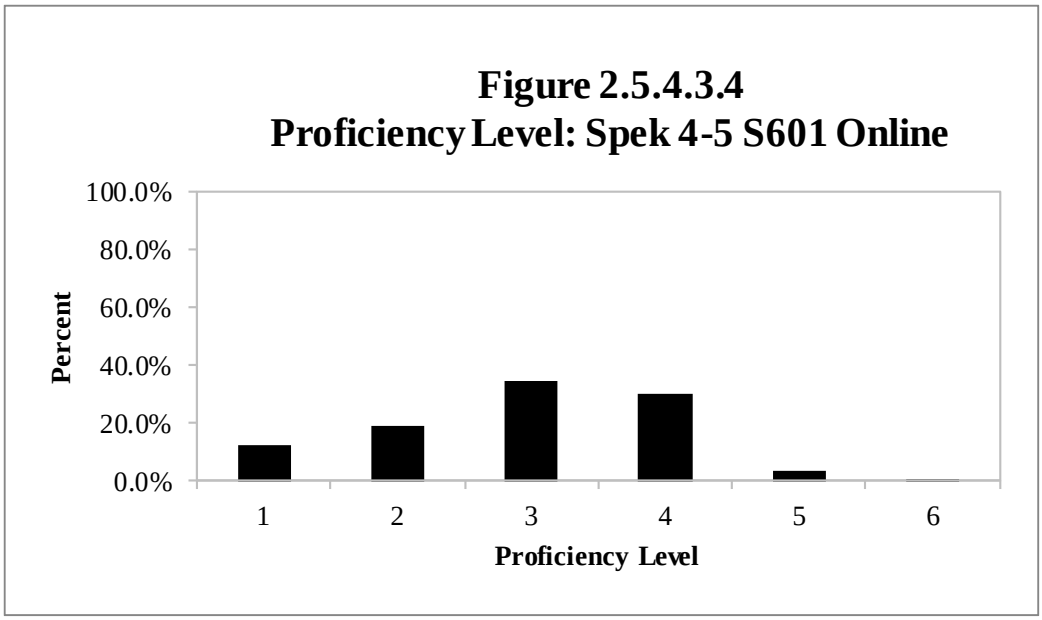


Table 2.5.4.3.4

Proficiency Level Distribution: Spek 4-5 S601 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	19,398	9.81%	25,649	15.64%	45,047	12.45%
2	35,483	17.94%	33,043	20.15%	68,526	18.94%
3	68,087	34.43%	56,374	34.37%	124,461	34.40%
4	64,619	32.68%	43,732	26.66%	108,351	29.95%
5	9,210	4.66%	4,968	3.03%	14,178	3.92%
6	963	0.49%	258	0.16%	1,221	0.34%
Total	197,760	100.00%	164,024	100.00%	361,784	100.00%



2.5.4.4 Grade 6-8

Table 2.5.4.4.1
Proficiency Level Distribution: Spek 6-8 Pre-A S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	2,916	100.00%	5,269	100.00%	8,313	100.00%	16,498	100.00%
Total	2,916	100.00%	5,269	100.00%	8,313	100.00%	16,498	100.00%

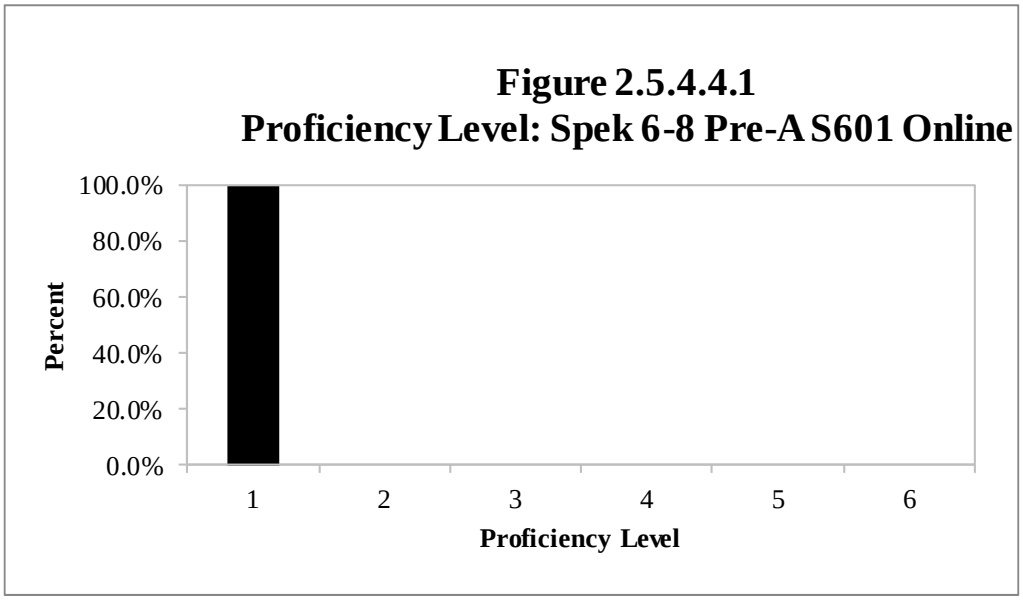


Table 2.5.4.4.2

Proficiency Level Distribution: Spek 6-8 A S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	16,674	57.37%	14,054	59.67%	17,328	39.30%	48,056	49.69%
2	9,170	31.55%	7,194	30.54%	16,668	37.81%	33,032	34.16%
3	2,862	9.85%	2,065	8.77%	8,395	19.04%	13,322	13.78%
4	346	1.19%	241	1.02%	1,684	3.82%	2,271	2.35%
5	13	0.04%	0	0.00%	14	0.03%	27	0.03%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	29,065	100.00%	23,554	100.00%	44,089	100.00%	96,708	100.00%

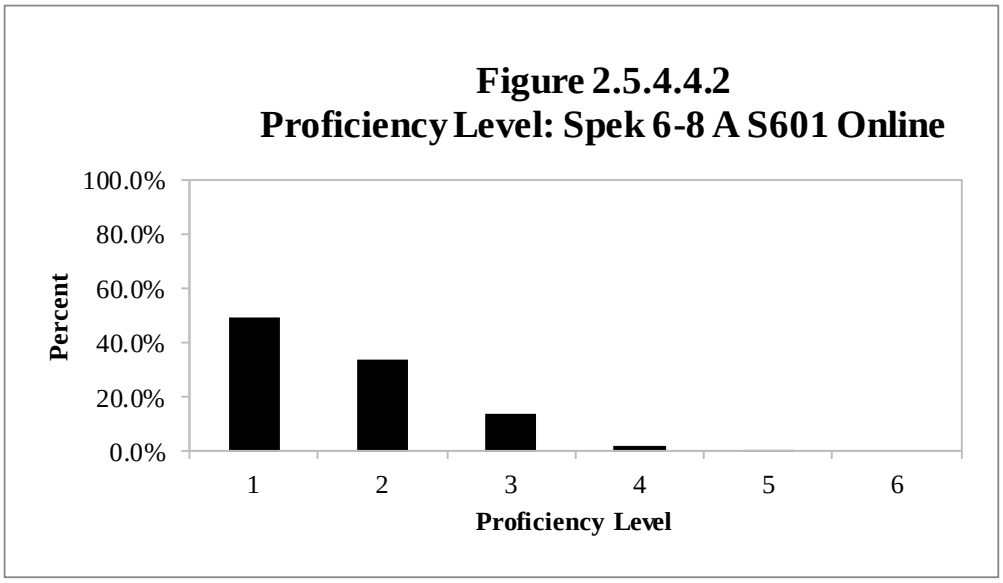


Table 2.5.4.4.3

Proficiency Level Distribution: Spek 6-8 B/C S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	9,552	8.90%	15,083	13.71%	9,654	12.05%	34,289	11.53%
2	28,406	26.47%	22,762	20.68%	17,284	21.58%	68,452	23.01%
3	47,895	44.63%	54,990	49.97%	34,233	42.74%	137,118	46.09%
4	20,578	19.17%	16,638	15.12%	18,625	23.25%	55,841	18.77%
5	881	0.82%	533	0.48%	240	0.30%	1,654	0.56%
6	14	0.01%	45	0.04%	64	0.08%	123	0.04%
Total	107,326	100.00%	110,051	100.00%	80,100	100.00%	297,477	100.00%

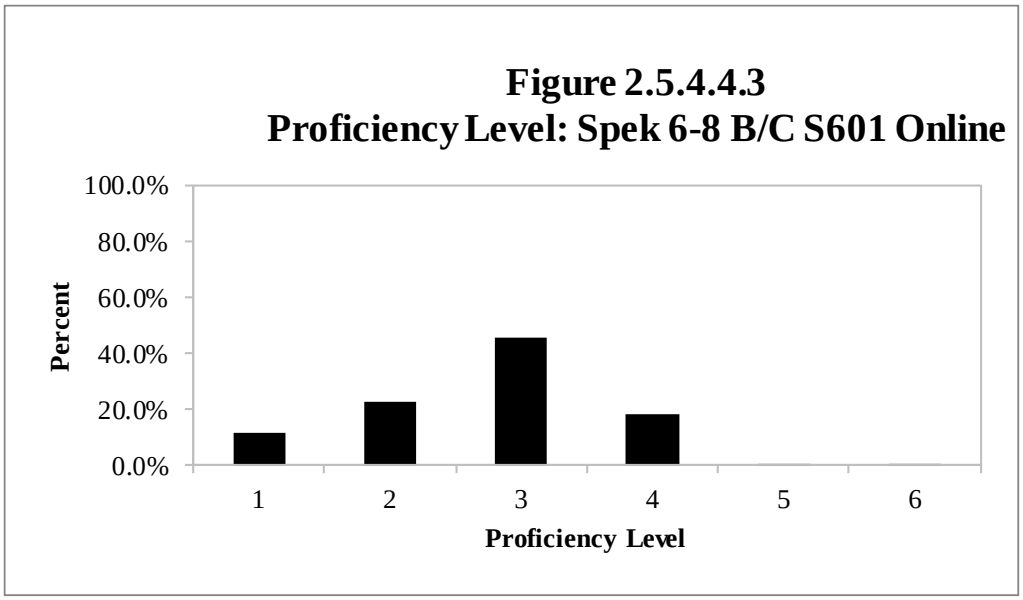
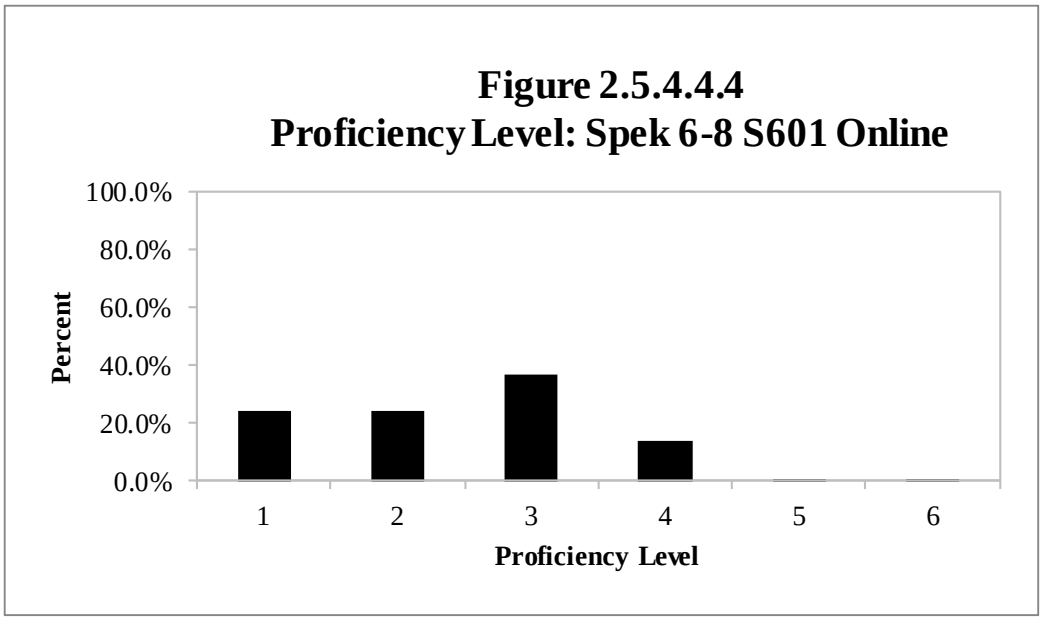


Table 2.5.4.4.4

Proficiency Level Distribution: Spek 6-8 S601 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	29,142	20.92%	34,406	24.77%	35,295	26.64%	98,843	24.07%
2	37,576	26.97%	29,956	21.57%	33,952	25.62%	101,484	24.71%
3	50,757	36.44%	57,055	41.08%	42,628	32.17%	150,440	36.63%
4	20,924	15.02%	16,879	12.15%	20,309	15.33%	58,112	14.15%
5	894	0.64%	533	0.38%	254	0.19%	1,681	0.41%
6	14	0.01%	45	0.03%	64	0.05%	123	0.03%
Total	139,307	100.00%	138,874	100.00%	132,502	100.00%	410,683	100.00%



2.5.4.5 Grade 9-12

Table 2.5.4.5.1
Proficiency Level Distribution: Spek 9-12 Pre-A S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	8,881	100.00%	9,217	100.00%	8,195	100.00%	7,614	100.00%	33,907	100.00%
Total	8,881	100.00%	9,217	100.00%	8,195	100.00%	7,614	100.00%	33,907	100.00%

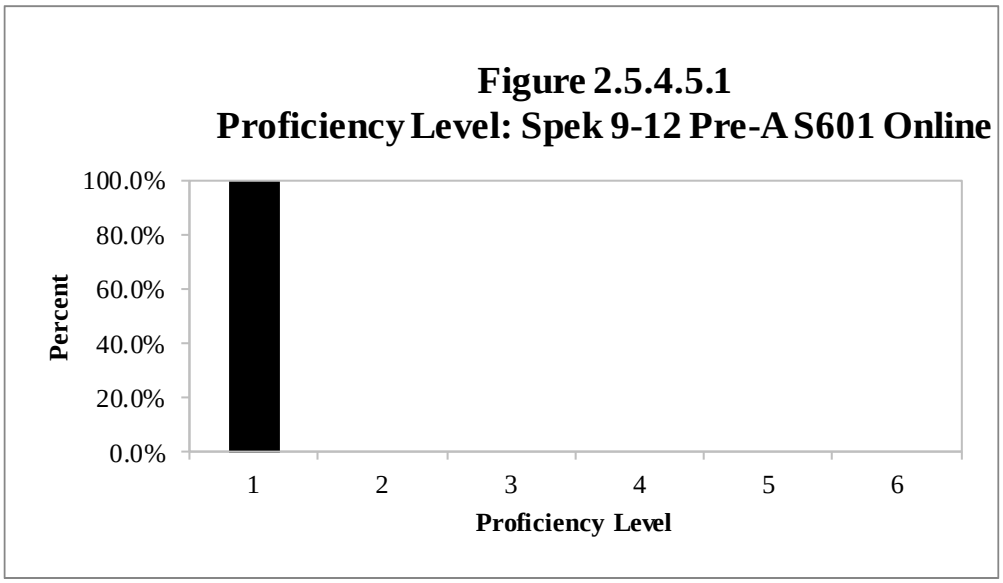


Table 2.5.4.5.2

Proficiency Level Distribution: Spek 9-12 A S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	51,850	60.82%	32,017	61.59%	12,206	64.44%	14,350	42.76%	110,423	58.20%
2	15,098	17.71%	9,318	17.93%	3,376	17.82%	12,753	38.00%	40,545	21.37%
3	17,122	20.08%	9,987	19.21%	3,179	16.78%	6,144	18.31%	36,432	19.20%
4	1,148	1.35%	658	1.27%	181	0.96%	310	0.92%	2,297	1.21%
5	39	0.05%	0	0.00%	0	0.00%	0	0.00%	39	0.02%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	85,257	100.00%	51,980	100.00%	18,942	100.00%	33,557	100.00%	189,736	100.00%

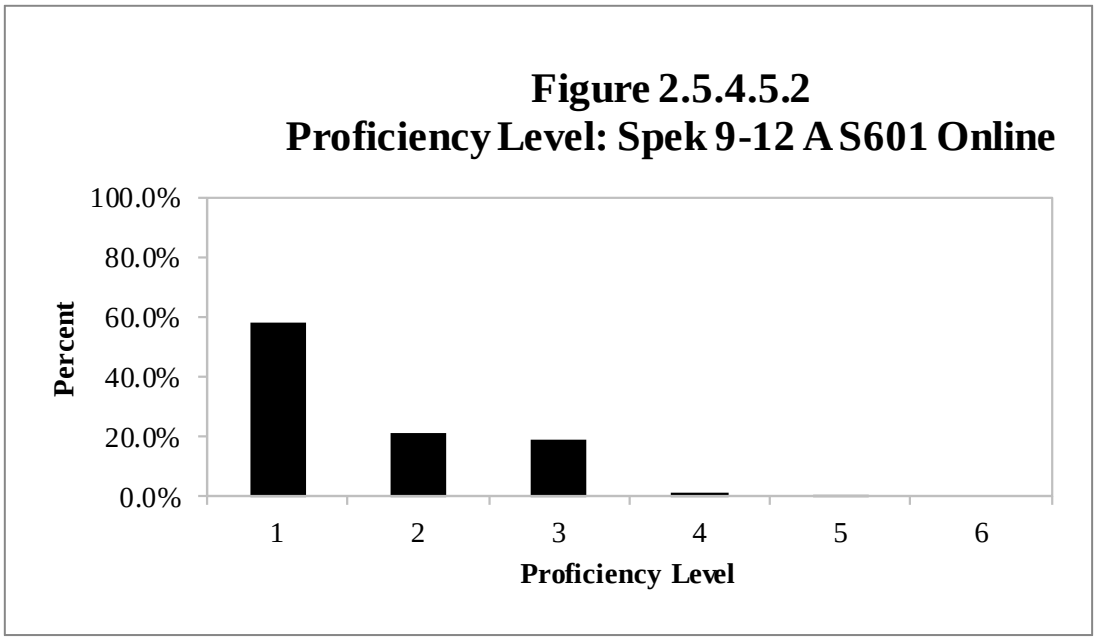


Table 2.5.4.5.3

Proficiency Level Distribution: Spek 9-12 B/C S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	7,553	12.65%	8,425	13.31%	13,037	21.12%	4,639	14.05%	33,654	15.46%
2	15,039	25.19%	14,915	23.56%	17,690	28.66%	8,285	25.09%	55,929	25.69%
3	30,298	50.76%	35,365	55.86%	26,126	42.33%	18,117	54.87%	109,906	50.48%
4	6,627	11.10%	4,407	6.96%	4,615	7.48%	1,803	5.46%	17,452	8.01%
5	117	0.20%	140	0.22%	182	0.29%	126	0.38%	565	0.26%
6	58	0.10%	58	0.09%	71	0.12%	50	0.15%	237	0.11%
Total	59,692	100.00%	63,310	100.00%	61,721	100.00%	33,020	100.00%	217,743	100.00%

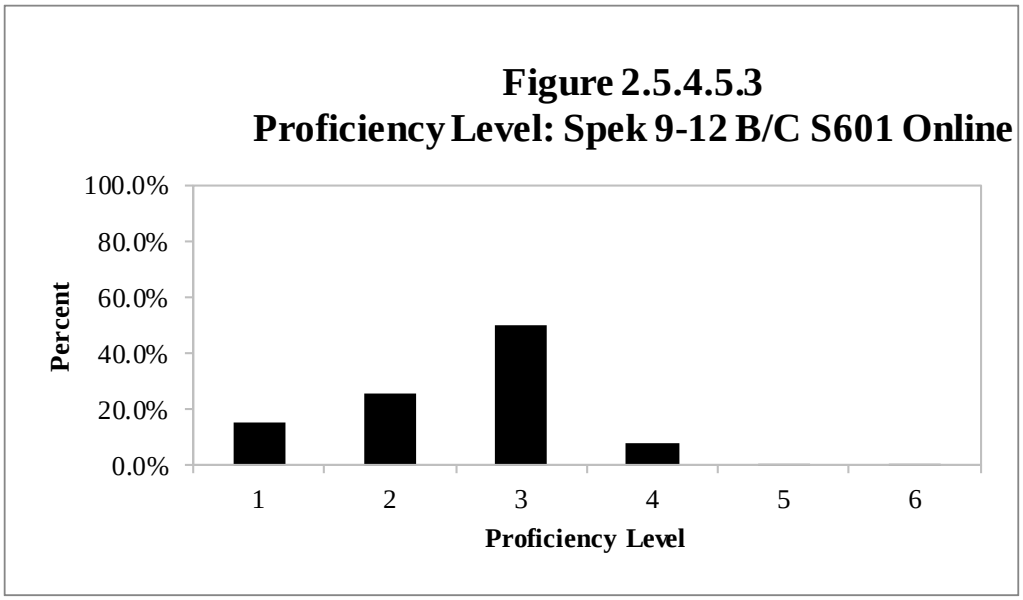
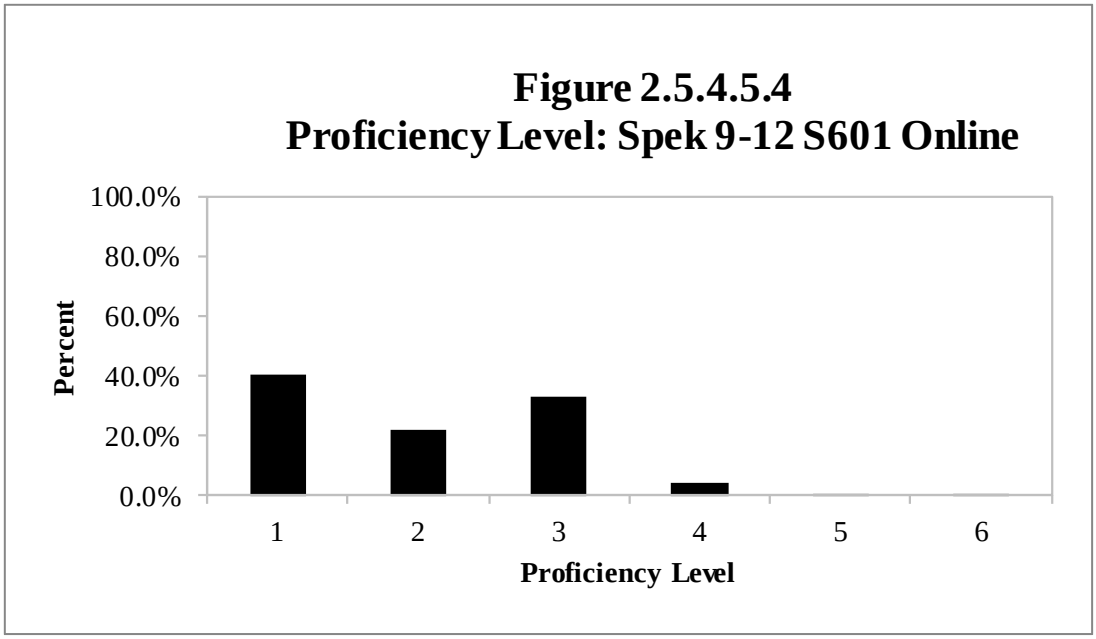


Table 2.5.4.5.4

Proficiency Level Distribution: Spek 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	68,284	44.39%	49,659	39.88%	33,438	37.63%	26,603	35.86%	177,984	40.32%
2	30,137	19.59%	24,233	19.46%	21,066	23.71%	21,038	28.36%	96,474	21.86%
3	47,420	30.83%	45,352	36.43%	29,305	32.98%	24,261	32.70%	146,338	33.15%
4	7,775	5.05%	5,065	4.07%	4,796	5.40%	2,113	2.85%	19,749	4.47%
5	156	0.10%	140	0.11%	182	0.20%	126	0.17%	604	0.14%
6	58	0.04%	58	0.05%	71	0.08%	50	0.07%	237	0.05%
Total	153,830	100.00%	124,507	100.00%	88,858	100.00%	74,191	100.00%	441,386	100.00%



2.6 Raw Score to Scale Score to Proficiency Level Conversion for Speaking and Writing

This section presents raw score to scale score conversions and associated proficiency levels for the test forms for Speaking and Writing.

The first column in the tables shows all possible raw scores. The second column shows the corresponding scale score. The third column shows the conditional standard error of measurement (CSEM) in the metric of the scale score, multiplied by 1.96. The resulting number (CSEM x 1.96) is used to construct the confidence band as reported on students' score reports. For example, if a student receives a scale score of 199 and if the CSEM multiplied by 1.96 is 45, then there is a 95% chance that the student's true scale score will be found somewhere between 154-244. For additional detail on conditional standard error of measurement, see Section 5, Reliability. Following the CSEM, columns provide the proficiency level interpretation for each grade in the grade-level cluster.

Performances that gain very few score points, and performances from students who gain all or almost all the score points, will have high CSEM values. The model does not precisely estimate these students' abilities; they may be well below or well above the range that is measured by the test and therefore the error of measurement is large. We provide further detail on the CSEM as it relates to the interpretation of student performances in Section 5.3, which provides CSEM values for proficiency level cuts.

Note that we truncate raw scores of zero where necessary, so that the lowest scale score given is the scale score corresponding to a proficiency level score of 1.0.

2.6.1 Listening

The ACCESS Online Listening test is a multistage adaptive assessment. As students do not all take the same set of items in the test, raw to scale score conversion tables are not presented.

2.6.2 Reading

The ACCESS Online Reading test is a multistage adaptive assessment. As students do not all take the same set of items in the test, raw to scale score conversion tables are not presented.

2.6.3 Writing

2.6.3.1 Grade 1

Table 2.6.3.1.1

Raw Score to Scale Score to Proficiency Level Conversion: Writ 1 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	111	256	1.0
1	194	45	1.6
2	208	33	1.7
3	217	29	1.8
4	225	28	1.8
5	233	28	1.9
6	241	31	2.0
7	252	35	2.3
8	265	39	2.7
9	281	41	3.0
10	299	42	3.3
11	316	42	3.6
12	333	40	3.9
13	349	38	4.2
14	362	36	4.5
15	375	36	4.8
16	389	40	5.3
17	409	52	6.0
18	441	94	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.3.1.2

Raw Score to Scale Score to Proficiency Level Conversion: Writ 1 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	111	256	1.0
1	200	45	1.7
2	214	32	1.8
3	223	28	1.8
4	230	27	1.9
5	238	28	2.0
6	247	31	2.2
7	257	35	2.5
8	271	39	2.8
9	287	41	3.1
10	304	42	3.4
11	322	42	3.7
12	339	40	4.0
13	354	38	4.3
14	368	36	4.6
15	381	36	4.9
16	395	40	5.5
17	415	52	6.0
18	446	94	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.3.2 Grades 2-3

Table 2.6.3.2.1

Raw Score to Scale Score to Proficiency Level Conversion: Writ 2-3 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	133	256	1.0	1.0
1	204	45	1.6	1.6
2	218	32	1.8	1.7
3	227	28	1.8	1.8
4	234	27	1.9	1.8
5	242	28	2.0	1.9
6	251	31	2.2	2.1
7	261	35	2.5	2.3
8	275	39	2.8	2.7
9	291	41	3.1	3.1
10	308	42	3.4	3.3
11	326	42	3.7	3.6
12	343	40	4.0	3.9
13	358	38	4.3	4.2
14	372	36	4.6	4.5
15	385	36	4.9	4.8
16	399	40	5.4	5.2
17	418	52	6.0	6.0
18	450	94	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.3.2.2

Raw Score to Scale Score to Proficiency Level Conversion: Writ 2-3 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	133	256	1.0	1.0
1	213	45	1.7	1.7
2	227	32	1.8	1.8
3	236	28	1.9	1.9
4	243	27	2.0	1.9
5	251	28	2.2	2.1
6	260	31	2.4	2.3
7	270	35	2.7	2.6
8	284	39	3.0	3.0
9	300	41	3.3	3.2
10	317	42	3.6	3.5
11	335	42	3.9	3.8
12	352	40	4.2	4.1
13	367	38	4.5	4.4
14	381	36	4.8	4.7
15	394	36	5.2	5.0
16	408	40	5.8	5.5
17	427	52	6.0	6.0
18	459	94	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.3.3 Grades 4-5

Table 2.6.3.3.1

Raw Score to Scale Score to Proficiency Level Conversion: Writ 4-5 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	155	256	1.0	1.0
1	235	45	1.7	1.7
2	249	32	1.8	1.8
3	258	28	1.9	1.9
4	266	27	2.0	1.9
5	274	28	2.3	2.2
6	282	31	2.7	2.5
7	293	35	3.0	3.0
8	306	39	3.2	3.2
9	322	41	3.5	3.4
10	340	42	3.8	3.7
11	358	42	4.1	4.0
12	375	40	4.4	4.3
13	390	38	4.7	4.6
14	403	36	5.0	4.9
15	416	36	5.6	5.3
16	430	40	6.0	5.8
17	450	52	6.0	6.0
18	482	94	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.3.3.2

Raw Score to Scale Score to Proficiency Level Conversion: Writ 4-5 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	155	256	1.0	1.0
1	264	45	1.9	1.9
2	278	33	2.5	2.4
3	288	29	3.0	2.8
4	296	28	3.1	3.0
5	304	29	3.2	3.1
6	313	31	3.3	3.3
7	323	34	3.5	3.4
8	337	38	3.7	3.6
9	353	41	4.0	3.9
10	370	42	4.3	4.2
11	388	42	4.7	4.6
12	405	40	5.1	4.9
13	420	38	5.7	5.5
14	433	36	6.0	6.0
15	447	37	6.0	6.0
16	461	40	6.0	6.0
17	481	52	6.0	6.0
18	513	94	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.3.4 Grades 6-8

Table 2.6.3.4.1

Raw Score to Scale Score to Proficiency Level Conversion: Writ 6-8 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	188	117	1.2	1.1	1.0
1	225	45	1.5	1.5	1.3
2	239	32	1.7	1.6	1.5
3	248	29	1.8	1.7	1.6
4	256	28	1.8	1.8	1.7
5	264	28	1.9	1.9	1.8
6	272	31	2.1	1.9	1.9
7	283	35	2.5	2.3	2.0
8	296	39	2.9	2.7	2.5
9	312	41	3.2	3.1	3.0
10	330	42	3.5	3.4	3.3
11	348	42	3.7	3.6	3.6
12	365	40	4.0	3.9	3.8
13	380	38	4.3	4.2	4.1
14	393	36	4.6	4.5	4.4
15	406	36	4.8	4.7	4.6
16	421	40	5.2	5.0	4.9
17	440	52	5.9	5.6	5.4
18	472	94	6.0	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.3.4.2

Raw Score to Scale Score to Proficiency Level Conversion: Writ 6-8 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	188	157	1.2	1.1	1.0
1	238	45	1.7	1.6	1.5
2	252	33	1.8	1.7	1.6
3	262	29	1.9	1.8	1.7
4	270	28	2.0	1.9	1.8
5	277	28	2.3	2.1	1.9
6	286	31	2.6	2.4	2.1
7	297	35	2.9	2.7	2.5
8	310	39	3.1	3.0	2.9
9	326	41	3.4	3.3	3.2
10	343	42	3.7	3.6	3.5
11	361	42	4.0	3.9	3.8
12	378	40	4.3	4.2	4.1
13	393	38	4.6	4.5	4.4
14	407	36	4.8	4.7	4.6
15	420	36	5.2	5.0	4.9
16	434	40	5.7	5.4	5.2
17	454	52	6.0	6.0	5.8
18	486	94	6.0	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.3.5 Grades 9-12

Table 2.6.3.5.1

Raw Score to Scale Score to Proficiency Level Conversion: Writ 9-12 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	232	99	1.3	1.2	1.1	1.0
1	262	45	1.7	1.5	1.4	1.3
2	276	32	1.8	1.7	1.6	1.5
3	285	28	1.9	1.8	1.7	1.6
4	293	27	2.1	1.9	1.8	1.7
5	300	28	2.3	2.0	1.9	1.7
6	309	31	2.6	2.3	2.0	1.8
7	319	35	3.0	2.7	2.4	2.0
8	333	39	3.2	3.1	2.9	2.5
9	349	41	3.5	3.3	3.2	3.0
10	366	42	3.7	3.6	3.5	3.4
11	384	42	4.1	3.9	3.8	3.7
12	401	40	4.4	4.3	4.2	4.0
13	416	38	4.7	4.6	4.5	4.3
14	430	36	5.0	4.8	4.7	4.6
15	443	36	5.3	5.1	5.0	4.9
16	457	40	5.6	5.4	5.3	5.1
17	477	52	6.0	5.9	5.7	5.5
18	508	94	6.0	6.0	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.3.5.2

Raw Score to Scale Score to Proficiency Level Conversion: Writ 9-12 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	232	87	1.3	1.2	1.1	1.0
1	257	45	1.6	1.5	1.4	1.2
2	271	33	1.8	1.6	1.5	1.4
3	281	30	1.9	1.8	1.6	1.5
4	289	28	2.0	1.8	1.7	1.6
5	298	29	2.3	2.0	1.8	1.7
6	307	31	2.6	2.3	1.9	1.8
7	317	34	2.9	2.6	2.3	1.9
8	330	38	3.1	3.0	2.8	2.4
9	346	41	3.4	3.3	3.1	3.0
10	363	42	3.7	3.6	3.5	3.3
11	381	42	4.0	3.9	3.8	3.6
12	398	40	4.3	4.2	4.1	4.0
13	413	38	4.6	4.5	4.4	4.3
14	427	37	4.9	4.8	4.7	4.5
15	440	37	5.2	5.0	4.9	4.8
16	455	40	5.6	5.4	5.2	5.1
17	475	52	6.0	5.9	5.6	5.5
18	506	94	6.0	6.0	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.4 Speaking

2.6.4.1 Grade 1

Table 2.6.4.1.1

Raw Score to Scale Score to Proficiency Level Conversion: Spek 1 Pre-A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	106	44	1.0
1	106	44	1.0
2	114	40	1.0
3	128	37	1.2
4	141	40	1.3
5	154	48	1.4
6	167	61	1.6

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.1.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 1 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	106	41	1.0
1	106	41	1.0
2	110	39	1.0
3	122	34	1.1
4	132	33	1.2
5	142	33	1.3
6	153	35	1.4
7	164	37	1.5
8	177	38	1.7
9	191	40	1.8
10	207	43	2.0
11	225	48	2.3
12	250	54	2.8
13	276	52	3.3
14	298	48	3.7
15	318	47	4.1
16	339	50	4.5
17	360	59	4.9
18	381	75	5.4

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.1.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 1 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
6	106	38	1.0
7	153	30	1.4
8	161	30	1.5
9	170	30	1.6
10	178	30	1.7
11	186	30	1.8
12	194	30	1.8
13	202	30	1.9
14	211	31	2.1
15	220	33	2.2
16	230	35	2.4
17	242	37	2.6
18	255	38	2.8
19	268	38	3.1
20	281	37	3.4
21	292	35	3.6
22	303	34	3.8
23	313	33	4.0
24	324	33	4.2
25	334	34	4.4
26	345	36	4.6
27	358	39	4.9
28	371	44	5.2
29	384	51	5.5
30	403	65	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.4.2 Grades 2-3

Table 2.6.4.2.1

Raw Score to Scale Score to Proficiency Level Conversion: Spek 2-3 Pre-A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	118	38	1.0	1.0
1	118	38	1.0	1.0
2	118	38	1.0	1.0
3	126	37	1.1	1.0
4	140	40	1.2	1.1
5	154	49	1.3	1.3
6	168	64	1.5	1.4

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.2.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 2-3 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	118	36	1.0	1.0
1	118	36	1.0	1.0
2	118	36	1.0	1.0
3	122	35	1.0	1.0
4	133	35	1.1	1.1
5	145	36	1.3	1.2
6	157	39	1.4	1.3
7	172	41	1.5	1.4
8	188	41	1.7	1.6
9	203	41	1.8	1.7
10	220	44	2.0	1.8
11	239	49	2.3	2.1
12	263	54	2.8	2.5
13	289	52	3.3	3.1
14	312	48	3.7	3.5
15	332	47	4.1	4.0
16	353	50	4.5	4.3
17	374	59	5.0	4.7
18	395	75	5.5	5.2

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.2.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 2-3 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
6	118	34	1.0	1.0
7	160	34	1.4	1.3
8	171	34	1.5	1.4
9	181	32	1.6	1.5
10	190	31	1.7	1.6
11	199	30	1.8	1.6
12	207	30	1.8	1.7
13	215	30	1.9	1.8
14	224	31	2.0	1.9
15	233	32	2.2	1.9
16	243	35	2.4	2.1
17	255	37	2.6	2.4
18	268	38	2.9	2.6
19	281	38	3.1	2.9
20	294	37	3.4	3.2
21	306	35	3.6	3.4
22	317	34	3.8	3.6
23	327	33	4.0	3.8
24	337	33	4.2	4.0
25	347	34	4.4	4.2
26	358	35	4.6	4.4
27	371	39	4.9	4.7
28	384	44	5.2	4.9
29	397	51	5.5	5.2
30	425	75	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.4.3 Grades 4-5

Table 2.6.4.3.1

Raw Score to Scale Score to Proficiency Level Conversion: Spek 4-5 Pre-A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	130	44	1.0	1.0
1	130	44	1.0	1.0
2	138	40	1.1	1.0
3	152	37	1.2	1.1
4	165	40	1.3	1.2
5	178	48	1.4	1.3
6	191	61	1.5	1.4

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.3.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 4-5 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	130	41	1.0	1.0
1	130	41	1.0	1.0
2	133	39	1.0	1.0
3	146	36	1.1	1.1
4	157	36	1.2	1.2
5	170	39	1.3	1.3
6	185	44	1.5	1.4
7	203	45	1.6	1.5
8	221	42	1.7	1.7
9	237	42	1.9	1.8
10	253	43	2.1	1.9
11	272	49	2.5	2.3
12	297	55	3.0	2.8
13	324	52	3.6	3.4
14	346	48	4.0	3.9
15	366	47	4.4	4.2
16	387	50	4.8	4.6
17	408	59	5.2	5.0
18	429	75	5.8	5.6

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.3.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 4-5 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
6	130	40	1.0	1.0
7	195	39	1.5	1.5
8	208	37	1.6	1.6
9	219	34	1.7	1.6
10	229	32	1.8	1.7
11	239	31	1.9	1.8
12	247	31	2.0	1.9
13	256	31	2.2	1.9
14	265	31	2.4	2.1
15	274	33	2.5	2.3
16	285	35	2.8	2.6
17	296	37	3.0	2.8
18	309	38	3.3	3.1
19	322	38	3.5	3.4
20	335	37	3.8	3.6
21	347	35	4.0	3.9
22	358	34	4.2	4.1
23	368	33	4.4	4.3
24	378	33	4.6	4.4
25	389	34	4.8	4.6
26	400	36	5.0	4.8
27	412	39	5.3	5.1
28	424	43	5.7	5.4
29	436	49	6.0	5.8
30	448	57	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.4.4 Grades 6-8

Table 2.6.4.4.1

Raw Score to Scale Score to Proficiency Level Conversion: Spek 6-8 Pre-A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	148	50	1.0	1.0	1.0
1	148	50	1.0	1.0	1.0
2	164	41	1.2	1.1	1.1
3	178	38	1.3	1.2	1.2
4	192	41	1.4	1.3	1.3
5	206	49	1.5	1.4	1.4
6	220	64	1.6	1.5	1.5

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.4.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 6-8 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	148	42	1.0	1.0	1.0
1	148	42	1.0	1.0	1.0
2	153	39	1.1	1.0	1.0
3	166	35	1.2	1.1	1.1
4	177	35	1.3	1.2	1.2
5	189	37	1.4	1.3	1.3
6	202	40	1.5	1.4	1.3
7	218	42	1.6	1.5	1.5
8	234	41	1.7	1.6	1.6
9	249	41	1.8	1.7	1.7
10	265	43	1.9	1.9	1.8
11	284	49	2.3	2.1	2.0
12	309	55	2.9	2.8	2.6
13	336	52	3.5	3.3	3.2
14	358	48	3.9	3.7	3.6
15	378	46	4.3	4.1	4.0
16	399	50	4.6	4.5	4.3
17	420	59	5.0	4.9	4.7
18	441	75	5.7	5.5	5.2

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.4.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 6-8 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
6	148	41	1.0	1.0	1.0
7	208	36	1.5	1.4	1.4
8	219	35	1.6	1.5	1.5
9	230	33	1.7	1.6	1.6
10	239	31	1.7	1.7	1.6
11	248	30	1.8	1.7	1.7
12	256	30	1.9	1.8	1.7
13	264	30	1.9	1.9	1.8
14	273	31	2.1	1.9	1.9
15	282	32	2.3	2.1	1.9
16	293	35	2.5	2.4	2.2
17	304	37	2.8	2.6	2.5
18	317	39	3.1	3.0	2.8
19	331	38	3.4	3.2	3.1
20	344	37	3.6	3.5	3.3
21	356	35	3.9	3.7	3.6
22	367	34	4.1	3.9	3.8
23	377	33	4.2	4.1	4.0
24	387	33	4.4	4.3	4.1
25	397	34	4.6	4.5	4.3
26	408	35	4.8	4.6	4.5
27	420	38	5.0	4.9	4.7
28	432	43	5.4	5.2	4.9
29	444	49	5.7	5.5	5.3
30	463	63	6.0	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.6.4.5 Grades 9-12

Table 2.6.4.5.1

Raw Score to Scale Score to Proficiency Level Conversion: Spek 9-12 Pre-A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	172	39	1.1	1.0	1.0	1.0
1	172	39	1.1	1.0	1.0	1.0
2	172	39	1.1	1.0	1.0	1.0
3	183	37	1.2	1.1	1.1	1.0
4	196	40	1.3	1.2	1.2	1.1
5	209	48	1.4	1.3	1.3	1.2
6	222	61	1.5	1.4	1.4	1.3

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.5.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 9-12 A S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	172	36	1.1	1.0	1.0	1.0
1	172	36	1.1	1.0	1.0	1.0
2	172	36	1.1	1.0	1.0	1.0
3	174	35	1.1	1.1	1.0	1.0
4	185	34	1.2	1.1	1.1	1.1
5	196	35	1.3	1.2	1.2	1.1
6	208	38	1.3	1.3	1.3	1.2
7	222	39	1.5	1.4	1.4	1.3
8	236	40	1.6	1.5	1.5	1.4
9	251	40	1.7	1.6	1.6	1.6
10	266	43	1.8	1.7	1.7	1.7
11	285	49	1.9	1.9	1.8	1.8
12	310	55	2.5	2.3	2.2	2.2
13	337	52	3.1	3.0	3.0	2.9
14	359	48	3.5	3.4	3.3	3.2
15	379	47	3.8	3.7	3.6	3.5
16	400	50	4.2	4.1	4.0	3.9
17	421	59	4.6	4.5	4.4	4.3
18	442	76	5.0	4.9	4.8	4.7

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

Table 2.6.4.5.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 9-12 B/C S601 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
6	172	34	1.1	1.0	1.0	1.0
7	213	34	1.4	1.3	1.3	1.3
8	223	33	1.5	1.4	1.4	1.3
9	233	32	1.5	1.5	1.5	1.4
10	242	31	1.6	1.6	1.5	1.5
11	250	30	1.7	1.6	1.6	1.6
12	259	30	1.7	1.7	1.6	1.6
13	267	30	1.8	1.7	1.7	1.7
14	276	31	1.8	1.8	1.8	1.8
15	285	33	1.9	1.9	1.8	1.8
16	296	35	2.1	2.0	1.9	1.9
17	307	37	2.4	2.3	2.2	2.1
18	320	38	2.7	2.6	2.5	2.4
19	334	38	3.1	3.0	2.9	2.8
20	346	37	3.3	3.2	3.1	3.0
21	358	35	3.5	3.4	3.3	3.2
22	369	34	3.7	3.6	3.5	3.4
23	379	33	3.8	3.7	3.6	3.5
24	389	33	4.0	3.9	3.8	3.7
25	400	34	4.2	4.1	4.0	3.9
26	411	36	4.4	4.3	4.2	4.1
27	423	39	4.6	4.5	4.4	4.3
28	435	43	4.9	4.7	4.6	4.5
29	455	55	5.5	5.3	5.1	5.0
30	476	72	6.0	6.0	6.0	6.0

Note: Score reports provided to students include the CSEM value multiplied by 1.96.

2.7 Equating Summary

Each year a certain number of items and tasks on the ACCESS for ELLs Online test form are new, as determined by the refreshment plan for that series. For Series 601, we refreshed all four domains.

For the Listening and Reading domains, WIDA implements a multiyear targeted refreshment plan to optimize the multistage computerized adaptive item pools and to ensure that we do not use these folders in the pools too long, thus overexposing them. In spring of 2020, WIDA and CAL assessment experts reviewed the 503 Listening and Reading item pools and identified folders that they believed the team should refresh for Series 601, according to the targeted refreshment plan. To meet these Series 601 targets, DRC field tested 40 Listening folders and 67 Reading folders.

For the Writing and Speaking domains, which are shorter, performance based, and which have additional content and exposure considerations in terms of task refreshment, WIDA and CAL assessment experts created the refreshment plan three years earlier to ensure that the test development effort could accommodate the refreshment target set for each series.

The Writing test consists of two sets of operational tasks that target four of the five WIDA ELD Standards. The first set targets Standard 2: Language of Language Arts and Standard 5: Language of Social Studies. The second set targets Standard 3: Language of Mathematics and Standard 4: Language of Science. The test creators designed each set of operational tasks, as well as each set of anchor tasks, to measure student performance across the entire proficiency scale, from PL 1 to PL 6. We refresh one of the two sets each year, on an alternating schedule, so the two WIDA ELD Standards that the anchor tasks target alternate from year to year.

The Speaking test consists of three sets of operational tasks that target all five WIDA ELD Standards. The first set targets Standard 1: Social and Instructional Language. The second set targets Standard 2: Language of Language Arts and Standard 5: Language of Social Studies. The third set targets Standard 3: Language of Mathematics and Standard 4: Language of Science. The test creators designed each set of operational tasks, as well as each set of anchor tasks, to measure student performance across the entire proficiency scale, from PL 1 to PL 6. Generally, we refresh one (or two) of the three sets each year on a rotating schedule, so the two WIDA ELD

Standards that the anchor tasks target also rotate from year to year. This allows for the Speaking test to be of manageable length and still contain embedded field test tasks, in consideration of the seat time required of students to complete each Speaking performance task. We refreshed two panels, or six tasks, for Series 601.

When we consider the sets of anchor tasks for the Speaking and Writing tests, it is important to note the overall assessment construct when we further consider the distribution of anchor tasks. The overarching goal of ACCESS for ELLs Online is to measure academic English language proficiency of students in each of the four domains. WIDA measures English language proficiency using a 6-level scale, which is defined in the WIDA Performance Definitions for the receptive domains (Listening and Reading) and productive domains (Speaking and Writing). WIDA does not have performance definitions that define a proficiency scale for each of the WIDA Standard Statements, e.g., no performance definitions exist specifically for Social and Instructional Language or the Language of Math. Given that proficiency in the WIDA Standard Statements is not defined, ACCESS for ELLs does not measure proficiency in the WIDA Standard Statements, and thus WIDA does not report proficiency scores for students at the level of the WIDA Standard Statements (See Part 1 of this report). Therefore, it is not necessary for the anchor sets in Speaking and Writing to contain tasks that target all five of the WIDA Standard Statements. Rather, it is more important to ensure that each anchor task assesses the targeted proficiency levels so we can sufficiently claim that ACCESS for ELLs Online truly measures across the breadth of the proficiency scale.

We used an equating procedure, known as common item equating, to equate the results from the new item/task pool and forms to the older item/task pool and forms using the common items/tasks, which are items/tasks that appear in both Series 503 and 601 for all domains. The characteristics of the common items/tasks were kept the same between series, as were the wording, formatting, and other test characteristics such as graphics. Furthermore, common items/tasks appeared in the same item/task sequence position as they appeared in the previous test series. In this procedure, we kept constant across both pools and test forms the difficulty measures for the items and tasks included on both the new and the old forms. In this way, the test user may employ the same frame of reference when interpreting students' scores on the newer test forms.

For the Listening and Reading domains, we used a pre-equating design to conduct the annual equating using student data collected from the Series 601 embedded field test (See Part 1, section 2.3.2). This design allowed for Listening and Reading item parameters to be available for setting up the computer adaptive engine prior to operational administration. We included in the final analyses all the student data that was available at the time that we conducted these equating analyses. All common items between Series 601 and 503 are used as anchors and were maintained in that role if they met two criteria: (1) the item/task displayed adequate fit (i.e., item/task mean square infit and outfit measures were between -1.30 and 1.30, and (2) the item/task exhibited no C-level or CC-level DIF. Using these criteria, we did not need to remove any common items/tasks from the anchor sets for any of the Series 601 tests prior to conducting the equating analysis. Because we included all Series 503 operational items in the anchor set when conducting the annual equating, the content representation of the anchor set was not a concern.

For both the Writing and Speaking tests, DRC implemented an embedded field test design (See Part 1, Section 2.3.2).

For the annual equating of the Writing test, DRC drew random samples of students from among those who had already taken the Writing test at the time of the draw, according to WIDA's predetermined sampling plan. When implementing that sampling plan, DRC drew a fixed number of students by grade cluster and tiered forms, where the number of students drawn was proportional to the population means of the number of students across previous series for the grade cluster and tiered forms.

For the annual equating of the Speaking test, DRC drew random samples of students from among those who had already taken the Speaking test at the time of the draw. When implementing that sampling plan, DRC drew a fixed number of students by grade cluster and tiered forms, where the number of students drawn was proportional to the population means of the number of students across previous series for the grade clusters. We included in the final analysis all the student data that was available at the time when we conducted our annual equating analyses.

The standard equating procedure involves anchoring all items/tasks common to Series 601 item/task pools and forms to their Series 503 values in the equating run, while the items and tasks parameters for new items and tasks were estimated. This procedure places the parameters

of the new 601 items and tasks on the same scale as those of the 503 items and tasks. For the Listening, Reading, and Speaking domains, we examined the displacement statistics of the anchored item/task after the first equating run. If the displacement statistics for any items and tasks is greater than the pre-established thresholds set by WIDA described below, the anchored items or tasks parameters will be re-estimated until the displacement statistics for all anchored items and tasks are below the thresholds. The **displacement statistic** shows the difference between the difficulty value of the anchored item/task and what its difficulty value would have been had we not anchored it. Smaller displacement statistics indicate more consistency between the item's (or task's) difficulty value on the Series 601 test form and on the Series 503 test form. Typically, displacements of less than 0.5 logits are unlikely to have much impact on measurement in a test instrument (Linacre, n.d.). For Listening and Reading items and P3 and P5 Speaking tasks, if this value was large (i.e., above 0.30 or below -0.30), that item was unanchored in the final equating run (i.e., it was treated as if it were a new item). For the Speaking P1 tasks, we used a slightly different displacement criterion (above 0.50 or below -0.50) since anchored P1 tasks from the Speaking domain have been found to be less stable than items and tasks from the other domains. Specifically, the test creators designed the Speaking P1 tasks to be very easy and therefore we can expect most students (98% to 99%) to get the full two points. As a result, the item difficulties for these P1 tasks are susceptible to small sampling fluctuations. A slight change in the percentages of students getting the full two points, due to sampling fluctuation, tends to cause the task difficulty values to change such that the displacement statistics will be out of the -0.3 and 0.3 range. If we were to use the same displacement criterion as other tasks, task difficulties for the P1 tasks would need to be re-estimated each time a slightly different sample is used to estimate them. Therefore, we used a more conservative estimate (-0.5 to 0.5) to evaluate the displacement statistics for the Speaking P1 tasks in order to ensure the stability of the Speaking scale scores. Since the Writing test has only one task anchored, there are no displacement statistics to evaluate.

Because of an item exposure issue of the Speaking equating sample, WIDA requested a modification to the equating procedure for the Speaking test. Specially, three new tasks (Task ID: 19928, 19935, and 19013) were exposed during the time the data of the equating sample were collected. Due to the concern that the equating sample's responses to these three tasks might have been compromised, CAL fixed the parameters of these tasks to their field test values

instead of estimating them using the equating sample. For the rest of the anchored tasks, CAL evaluated their displacement statistics using normal procedure.

The tables that follow present a summary of the equating results. The first section of each table compares the current test (i.e., the Series 601 version of that item/task pool and test form) to the previous year's test (i.e., the Series 503 version of that item/task pool and test form). The table shows the number of items/tasks, the average item/task difficulty, the standard deviation of the item/task difficulty values, and the difficulty value of the easiest and hardest item/task on each test form. These values are in log-odd units, or **logits** (i.e., analyses carried out using Rasch measurement techniques, which produce equal-interval, linear measures expressed on a logit scale). In the domains of Listening and Reading, if the equating is successful, we would expect the average item difficulty values for the two series to be similar. This is true for these domains because they have many test items in the item pool, as well as large anchor sets. Additionally, the Series 601 Writing domain tests consist of only two tasks, with only one task serving as an anchor between series. Therefore, we might expect some differences in the average difficulty values for the two Writing series. Similarly, we might expect some differences in the average difficulty values for the two Speaking series, as those test forms included only nine tasks, and one-third of the test served as the anchor between series.

The second section of each table presents information about the anchor items/tasks and shows the total number of possible anchors that we initially anchored to the values from the previous series, as well as the average item/task difficulty and the average standard deviation of the difficulty values for those items/tasks. Next, the table shows the number of items/tasks that we anchored in the final equating run, again with the average item/task difficulty and the average standard deviation of those difficulty values for those items/tasks. Finally, the table gives the percentage of items/tasks that served as anchors and their average displacement values. In general, the larger the number and the higher the percentage of items/tasks anchored and the closer their average displacement is to 0.00, the more trustworthy the equating results will be (Johns & Smith, 2006; Stahl & Muckle, 2007).

The third section of each table gives information about the anchor items/tasks, both by order of displacement statistics and by order of item/task difficulty. The displacement statistics provide information regarding the difference between the difficulty value of each anchored item/task and

what that difficulty value would have been had we not anchored the item/task. Smaller displacement statistics indicate more consistency between the item's (or task's) difficulty value between the Series 601 test form and on the Series 503 test form. The anchor items/tasks appearing on a given test form should have a range of item/task difficulties that mirrors the range of item/task difficulties in the entire pool (Kolen & Brennen, 2004).

The tables for the Writing and Speaking domains have a fourth section, which provides the anchored **Rasch rating scale model step measures** for each task (also known as Rasch structure calibrations, step parameters, step calibrations, or Rasch-Andrich thresholds). Step measures identify the particular points along the student proficiency continuum where it is equally probable that a rater evaluating a student's response to a task would have assigned a score in either of two adjacent score categories. That is, a step measure indicates how likely it is for a student to receive a score in a particular score category relative to the adjacent score category on that scale. It is not a measure of the difficulty of the category (Linacre, 2004).

If the score categories are working as those who designed the scoring scale intended, the step measures should advance from step to step by at least 1.4 logits, but not more than 5.0 logits (Linacre, 2004). However, the required degree of advancement in the step measures lessens as the number of score categories increase. For practical purposes, advances of 1.4 logits are generally not required to be able to make valid inferences regarding a student's level of proficiency based on their score (Linacre, 2004).

If the step measures do not advance, then that indicates that the raters likely assigned few scores in one (or more) score categories, resulting in a set of "disordered" thresholds. When the frequency of scores that raters assigned in a category is low, then the step measure for that category will be imprecisely estimated and potentially unstable (Linacre, 2004).

For the Writing test forms, multiple tasks appeared on each form. We employed a rating scale model to analyze the scores that the raters assigned to students' written responses to those tasks. When using this model, we assumed that the raters used the score categories in a similar manner when assigning scores to students' responses to both tasks included on the test form. That is, under this assumption, when Winsteps analyzed the students' Writing scores, it treated the 3s that raters assigned to students' responses to one task as equivalent to the 3s that raters assigned to students' responses on another task. Similarly, the computer program treats the 4s that raters

assigned to students' responses to one task as equivalent to the 4s that raters assigned to students' responses on another task. Accordingly, the output from the Winsteps analysis reports a single set of step measures that applied to both the Writing tasks appearing on that test form. The Writing step measures advanced from step to step except from Step 1 to Step 2, which indicated that raters tended to assign fewer scores of 1 when compared with the other score categories. The advances in the step measures ranged from 0.17 logits (from Step 2 to Step 3) to 1.28 logits (from Step 6 to Step 7). While these findings do not signal optimal scoring scale functioning (i.e., the step measures did not advance from step to step by at least 1.4 logits), raters' use of the Writing Scoring Scale should still yield student scores that test users can meaningfully interpret (Linacre, 2004). To provide anchors for the calibration of new Writing tasks, to facilitate their placement onto the common WIDA score scale each year, we held the step measures constant.

For the Speaking test forms, we used a rating scale model to analyze the scores that raters assigned students' responses to all the PL 1 tasks, assuming that raters used the three score categories (0–2) on that scoring scale in a similar manner when evaluating students' oral responses to those tasks. Similarly, we used the same rating scale model to analyze the scores that raters assigned students' responses to the PL 3 and PL 5 tasks, assuming that raters used the five score categories (0-4) on that scoring scale in a similar manner when evaluating students' oral responses to those tasks. Therefore, the step measures for all PL 1 tasks were the same, and the step measures for all PL 3 and PL 5 tasks were the same. The Speaking step measures advanced from step to step for the PL 1 tasks and for the PL 3 and PL 5 tasks. For the PL 1 tasks, the step measures advanced by 1.12 logits from Step 1 to Step 2. For the PL 3 and PL5 tasks, the advances in the step measures ranged from 0.85 logits (from Step 1 to Step 2) to 3.26 logits (from Step 2 to Step 3). While these findings do not signal optimal scoring scale functioning (i.e., the step measures did not all advance from step to step by at least 1.4 logits), raters' use of the two Speaking Scoring Scales should still yield student scores that test users can meaningfully interpret (Linacre, 2004). As with Writing, these constant step measures help to provide anchors in the calibration of new Speaking tasks, facilitating their placement onto the common WIDA score scale each year.

The tables in the next section of this report reveal that the average difficulty levels for the items appearing on the Series 601 Listening and Reading test forms were similar to those for the previous series for all grade-level clusters. For the Listening domain, the differences in the

average difficulty levels ranged from 0.02 logits (for Grade 1 and Grades 6-8) to 0.07 logits (for Grades 2-3 and Grades 4-5). Similarly, for the Reading domain, the differences in the average difficulty levels ranged from 0.01 logits (for Grades 6-8) to 0.1 logits (for Grades 4-5). For each Listening and Reading test form, the anchor items represented a wide range of difficulties that spanned nearly the entire item difficulty continuum.

The differences in the average difficulty levels for the tasks appearing on the Writing test forms for Series 601 and 503 were less than 0.20 logits for all grade-level clusters and tiers, except for Grades 4–5Tier B/C. For Grades 4-5 Tier B/C, the difference was 0.25 logits.

The differences in the average difficulty levels for the tasks appearing on the Speaking test forms for Series 601 and 503 were less than 0.20 logits for all grade-level clusters except for Grades 2-3 and Grades 9-1. For Grades 2-3, the difference was 0.20 logits; for Grades 9-1, the difference was 0.21 logits. For each Speaking test form, the anchor tasks represented a range of difficulties that spanned nearly the entire task difficulty continuum.

WIDA psychometricians reviewed the equating plans before CAL conducted the equating analyses. The WIDA psychometricians then reviewed the equating results at the conclusion of the equating project to ensure that the equating was carried out correctly and the results were deemed reasonable. Besides the evidence listed above to the success of the equating results, WIDA and CAL psychometricians compare scoring tables across years to ensure that scores are comparable across test series, which demonstrates that the tests are comparable across series. In addition, WIDA and CAL psychometricians reviewed the annual equating results and identified issues that they felt they needed to bring to the attention of the WIDA Technical Advisory Committee.

2.7.1 Listening

2.7.1.1 Grade 1

Table 2.7.1.1

Equating Summary: List 1 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	54	-1.11 (1.06)	54	-1.09 (1.11)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-3.59	0.96	-3.59	1.13
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	45	-1.18 (1.08)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	45	-1.18 (1.08)		
	Percentage Anchors	**Average Displacement**		
	83%	0.02		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	18874	-0.58	-0.69	18875	-3.59	-0.05
	14898	-1.93	-0.38	14952	-3.03	0.05
	18873	-1.01	-0.33	18841	-3.01	0.21
	17814	-1.12	-0.26	13889	-2.96	0.19
	17781	-2.96	-0.15	17781	-2.96	-0.15
	16560	-0.02	-0.13	17779	-2.57	-0.11
	16559	0.50	-0.13	13891	-2.55	0.03
	17779	-2.57	-0.11	17813	-2.32	0.13
	18891	0.20	-0.11	13890	-2.23	0.20
	14897	-1.30	-0.10	18842	-2.16	0.47
	16558	-0.15	-0.08	14898	-1.93	-0.38
	16641	-0.86	-0.07	16531	-1.79	0.01
	17793	-0.27	-0.06	17815	-1.76	0.22
	16533	-0.47	-0.06	14951	-1.68	0.24
	17788	0.01	-0.06	13900	-1.63	0.04

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	18875	-3.59	-0.05	17791	-1.57	0.18
	16642	-0.74	-0.03	17780	-1.52	0.11
	18890	0.18	-0.03	14899	-1.50	0.25
	18843	-0.99	-0.03	14953	-1.43	0.24
	13899	-1.15	-0.02	14897	-1.30	-0.10
	16531	-1.79	0.01	16535	-1.22	0.06
	19330	0.21	0.02	13899	-1.15	-0.02
	13891	-2.55	0.03	17814	-1.12	-0.26
	13900	-1.63	0.04	18873	-1.01	-0.33
	14952	-3.03	0.05	18843	-0.99	-0.03
	16535	-1.22	0.06	13898	-0.94	0.06
	13898	-0.94	0.06	16641	-0.86	-0.07
	18889	-0.69	0.06	19514	-0.77	0.09
	19513	0.12	0.08	16642	-0.74	-0.03
	19514	-0.77	0.09	18889	-0.69	0.06
	17780	-1.52	0.11	18874	-0.58	-0.69
	17813	-2.32	0.13	16533	-0.47	-0.06
	16640	-0.25	0.16	19332	-0.32	0.20
	17791	-1.57	0.18	17793	-0.27	-0.06
	13889	-2.96	0.19	16640	-0.25	0.16
	19332	-0.32	0.20	19512	-0.19	0.24
	13890	-2.23	0.20	16558	-0.15	-0.08
	19331	0.87	0.21	16560	-0.02	-0.13
	18841	-3.01	0.21	17788	0.01	-0.06
	17815	-1.76	0.22	19513	0.12	0.08
	14953	-1.43	0.24	18890	0.18	-0.03
	19512	-0.19	0.24	18891	0.20	-0.11
	14951	-1.68	0.24	19330	0.21	0.02
	14899	-1.50	0.25	16559	0.50	-0.13
	18842	-2.16	0.47	19331	0.87	0.21

2.7.1.2 Grade 2-3

Table 2.7.1.2

Equating Summary: List 2-3 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	54	-0.84 (1.77)	54	-0.91 (1.80)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-4.25	2.60	-4.25	2.60
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	48	-0.87 (1.83)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	48	-0.87 (1.83)		
	Percentage Anchors	**Average Displacement**		
	89%	0.03		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	19367	0.69	-1.03	17769	-4.25	0.16
	19368	1.52	-0.96	13788	-4.17	0.22
	18893	1.44	-0.74	14879	-4.12	0.44
	18867	-1.89	-0.49	13905	-3.26	0.41
	18895	-0.93	-0.37	12825	-3.26	0.19
	16653	0.97	-0.37	13904	-3.24	0.30
	16654	1.15	-0.33	12956	-3.09	0.50
	16684	2.02	-0.30	18865	-3.03	-0.01
	19366	-1.11	-0.26	17770	-2.72	0.18
	13789	-0.72	-0.23	13790	-2.68	0.10
	18894	-1.65	-0.21	14884	-2.64	0.64
	16685	0.53	-0.19	16602	-2.56	0.03
	18866	-2.32	-0.18	13910	-2.33	0.27
	19342	2.60	-0.15	18866	-2.32	-0.18
	12828	-2.29	-0.07	12828	-2.29	-0.07

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	16652	-1.17	-0.02	13906	-2.24	0.20
	18865	-3.03	-0.01	18867	-1.89	-0.49
	16602	-2.56	0.03	18894	-1.65	-0.21
	17771	-1.07	0.05	12830	-1.17	0.31
	19350	0.81	0.07	16652	-1.17	-0.02
	16686	-0.47	0.08	12957	-1.12	0.38
	13790	-2.68	0.10	19366	-1.11	-0.26
	12971	0.35	0.11	17771	-1.07	0.05
	19492	0.51	0.13	18895	-0.93	-0.37
	19494	1.51	0.13	13789	-0.72	-0.23
	13912	-0.24	0.14	13911	-0.58	0.15
	13911	-0.58	0.15	16686	-0.47	0.08
	17769	-4.25	0.16	16603	-0.35	0.17
	16603	-0.35	0.17	13912	-0.24	0.14
	17770	-2.72	0.18	16604	-0.07	0.21
	12825	-3.26	0.19	19493	0.22	0.35
	19352	1.76	0.19	12971	0.35	0.11
	13906	-2.24	0.20	19344	0.37	0.32
	16604	-0.07	0.21	14883	0.51	0.33
	13788	-4.17	0.22	19492	0.51	0.13
	19343	0.97	0.22	16685	0.53	-0.19
	13910	-2.33	0.27	19367	0.69	-1.03
	19351	0.87	0.28	19350	0.81	0.07
	13904	-3.24	0.30	19351	0.87	0.28
	12830	-1.17	0.31	16653	0.97	-0.37
	19344	0.37	0.32	19343	0.97	0.22
	14883	0.51	0.33	16654	1.15	-0.33
	19493	0.22	0.35	18893	1.44	-0.74
	12957	-1.12	0.38	19494	1.51	0.13
	13905	-3.26	0.41	19368	1.52	-0.96
	14879	-4.12	0.44	19352	1.76	0.19
	12956	-3.09	0.50	16684	2.02	-0.30
	14884	-2.64	0.64	19342	2.60	-0.15

2.7.1.3 Grades 4-5

Table 2.7.1.3

Equating Summary: List 4-5 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	54	0.66 (1.37)	54	0.60 (1.45)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-2.36	3.33	-2.65	3.33
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	51	0.70 (1.40)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	51	0.70 (1.40)		
	Percentage Anchors	**Average Displacement**		
	94%	-0.04		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	18716	1.64	-0.88	12793	-2.36	-0.05
	18719	1.07	-0.88	16618	-2.08	0.54
	13028	3.20	-0.64	12792	-1.97	0.62
	18714	1.22	-0.54	13024	-1.78	0.01
	13027	3.33	-0.44	19521	-1.16	0.25
	14939	2.28	-0.44	16613	-0.78	0.24
	18715	1.29	-0.43	13026	-0.76	0.15
	16619	2.15	-0.36	16708	-0.72	0.19
	12794	-0.22	-0.32	18720	-0.52	-0.19
	16714	2.68	-0.21	16710	-0.52	0.11
	16620	2.50	-0.20	18628	-0.49	-0.01
	18720	-0.52	-0.19	19520	-0.24	0.19
	19424	2.59	-0.14	12794	-0.22	-0.32
	19425	2.18	-0.14	18718	-0.11	0.19
	17789	0.08	-0.14	13025	-0.05	0.00

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	14940	0.48	-0.14	14947	-0.04	0.37
	16709	1.19	-0.13	17789	0.08	-0.14
	14946	0.95	-0.13	14945	0.10	0.04
	13029	1.83	-0.12	16615	0.10	-0.10
	19426	2.11	-0.11	18629	0.24	0.55
	16713	1.64	-0.11	16616	0.29	0.06
	16615	0.10	-0.10	14941	0.32	0.10
	17792	0.70	-0.09	19522	0.38	0.12
	12793	-2.36	-0.05	18627	0.44	0.08
	18617	2.68	-0.02	17790	0.45	0.07
	18628	-0.49	-0.01	14940	0.48	-0.14
	13025	-0.05	0.00	16712	0.50	0.11
	13024	-1.78	0.01	17792	0.70	-0.09
	19372	2.32	0.03	14946	0.95	-0.13
	14945	0.10	0.04	18719	1.07	-0.88
	16616	0.29	0.06	16709	1.19	-0.13
	17790	0.45	0.07	19370	1.19	0.07
	19370	1.19	0.07	18616	1.21	0.22
	18627	0.44	0.08	18714	1.22	-0.54
	14941	0.32	0.10	18715	1.29	-0.43
	16710	-0.52	0.11	18615	1.58	0.56
	16712	0.50	0.11	16713	1.64	-0.11
	19522	0.38	0.12	18716	1.64	-0.88
	19371	2.64	0.13	13029	1.83	-0.12
	13026	-0.76	0.15	19426	2.11	-0.11
	18718	-0.11	0.19	16619	2.15	-0.36
	16708	-0.72	0.19	19425	2.18	-0.14
	19520	-0.24	0.19	14939	2.28	-0.44
	18616	1.21	0.22	19372	2.32	0.03
	16613	-0.78	0.24	16620	2.50	-0.20

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	19521	-1.16	0.25	19424	2.59	-0.14
	14947	-0.04	0.37	19371	2.64	0.13
	16618	-2.08	0.54	18617	2.68	-0.02
	18629	0.24	0.55	16714	2.68	-0.21
	18615	1.58	0.56	13028	3.20	-0.64
	12792	-1.97	0.62	13027	3.33	-0.44

2.7.1.4 Grades 6-8

Table 2.7.1.4

Equating Summary: List 6-8 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	54	1.19 (1.05)	54	1.21 (1.21)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-1.14	3.49	-1.14	3.49
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	42	1.19 (1.03)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	42	1.19 (1.03)		
	Percentage Anchors	**Average Displacement**		
	78%	-0.04		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	18588	1.29	-0.92	17679	-1.14	0.17
	18899	0.11	-0.55	18898	-0.23	0.16
	18586	1.09	-0.54	17678	-0.20	0.18
	18587	0.38	-0.52	19287	-0.06	0.00
	16563	0.06	-0.46	18897	0.00	0.21
	19288	1.93	-0.45	17680	0.05	0.00
	14863	1.43	-0.30	16563	0.06	-0.46
	16568	2.04	-0.22	16664	0.10	0.20
	19445	0.86	-0.17	18899	0.11	-0.55
	17727	3.33	-0.13	16562	0.24	0.21
	16566	1.86	-0.11	18587	0.38	-0.52
	14917	1.05	-0.08	19444	0.67	-0.03
	14916	1.73	-0.05	19286	0.71	0.04
	19446	1.12	-0.04	19445	0.86	-0.17
	19444	0.67	-0.03	17694	0.89	0.01

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	14859	1.51	-0.03	14917	1.05	-0.08
	17695	1.44	-0.02	17693	1.06	0.05
	16666	1.39	-0.01	17725	1.07	0.20
	19287	-0.06	0.00	14915	1.08	0.21
	16665	1.13	0.00	18586	1.09	-0.54
	17680	0.05	0.00	19446	1.12	-0.04
	17694	0.89	0.01	16665	1.13	0.00
	18600	3.49	0.04	18588	1.29	-0.92
	18599	3.27	0.04	19320	1.34	0.30
	19286	0.71	0.04	16666	1.39	-0.01
	17693	1.06	0.05	14863	1.43	-0.30
	14858	1.54	0.05	17695	1.44	-0.02
	18598	2.65	0.07	14859	1.51	-0.03
	19318	2.04	0.14	17726	1.54	0.14
	17726	1.54	0.14	14858	1.54	0.05
	18898	-0.23	0.16	19319	1.56	0.24
	16564	1.63	0.16	16564	1.63	0.16
	17679	-1.14	0.17	14916	1.73	-0.05
	17678	-0.20	0.18	16566	1.86	-0.11
	16664	0.10	0.20	19288	1.93	-0.45
	17725	1.07	0.20	16568	2.04	-0.22
	16562	0.24	0.21	19318	2.04	0.14
	18897	0.00	0.21	18598	2.65	0.07
	14915	1.08	0.21	16567	3.07	0.23
	16567	3.07	0.23	18599	3.27	0.04
	19319	1.56	0.24	17727	3.33	-0.13
	19320	1.34	0.30	18600	3.49	0.04

2.7.1.5 Grades 9-12

Table 2.7.1.5

Equating Summary: List 9-12 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	54	1.66 (1.10)	54	1.62 (1.13)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-0.48	4.08	-0.48	4.08
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	42	1.55 (1.15)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	42	1.55 (1.15)		
	Percentage Anchors	**Average Displacement**		
	78%	0.00		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	19304	0.85	-0.56	18573	-0.48	-0.01
	18575	1.70	-0.49	17713	-0.45	0.10
	19312	1.83	-0.47	17761	-0.38	0.04
	19303	1.36	-0.42	19310	-0.14	0.15
	17751	3.50	-0.28	17719	0.10	-0.05
	17755	4.08	-0.27	18574	0.26	0.24
	16658	2.48	-0.25	18566	0.28	0.03
	16656	2.18	-0.23	18565	0.53	0.26
	16587	2.22	-0.18	17741	0.54	0.07
	16586	1.08	-0.18	19302	0.77	0.06
	17749	2.88	-0.16	19311	0.79	0.16
	17754	2.67	-0.14	19304	0.85	-0.56
	17721	2.25	-0.13	16588	0.92	0.01
	17750	1.98	-0.12	17712	0.93	0.20
	16657	1.04	-0.09	17762	0.94	0.14

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	17743	1.77	-0.05	17715	0.99	0.15
	17719	0.10	-0.05	16657	1.04	-0.09
	17753	2.10	-0.02	16586	1.08	-0.18
	18573	-0.48	-0.01	17720	1.18	0.09
	16588	0.92	0.01	17742	1.34	0.16
	18566	0.28	0.03	19303	1.36	-0.42
	17761	-0.38	0.04	18567	1.48	0.32
	19290	2.86	0.05	17763	1.62	0.08
	19302	0.77	0.06	18575	1.70	-0.49
	17741	0.54	0.07	17743	1.77	-0.05
	17763	1.62	0.08	19312	1.83	-0.47
	17720	1.18	0.09	17750	1.98	-0.12
	17713	-0.45	0.10	17753	2.10	-0.02
	17762	0.94	0.14	16656	2.18	-0.23
	19292	2.93	0.14	16587	2.22	-0.18
	19310	-0.14	0.15	17721	2.25	-0.13
	17715	0.99	0.15	16658	2.48	-0.25
	17742	1.34	0.16	17754	2.67	-0.14
	19311	0.79	0.16	19358	2.70	0.26
	17712	0.93	0.20	19290	2.86	0.05
	18574	0.26	0.24	17749	2.88	-0.16
	19358	2.70	0.26	19291	2.90	0.48
	18565	0.53	0.26	19292	2.93	0.14
	18567	1.48	0.32	19360	3.00	0.47
	19359	3.35	0.36	19359	3.35	0.36
	19360	3.00	0.47	17751	3.50	-0.28
	19291	2.90	0.48	17755	4.08	-0.27

2.7.2 Reading

2.7.2.1 Grade 1

Table 2.7.2.1

Equating Summary: Read 1 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	72	-0.98 (0.99)	72	-1.04 (1.08)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-3.60	0.84	-4.07	0.84
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	57	-0.93 (0.97)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	57	-0.93 (0.97)		
	Percentage Anchors	**Average Displacement**		
	79%	-0.05		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	18542	-3.52	-0.74	18458	-3.60	-0.54
	18543	-1.39	-0.63	18542	-3.52	-0.74
	18463	-1.14	-0.58	18537	-2.88	0.29
	18458	-3.60	-0.54	17975	-2.45	0.08
	17988	0.79	-0.48	17974	-2.43	0.25
	17959	0.18	-0.38	17954	-2.20	-0.05
	18450	-0.95	-0.35	18467	-2.16	-0.17
	18098	0.02	-0.29	18465	-2.12	-0.07
	17135	-0.85	-0.25	13193	-2.11	0.21
	17982	-0.38	-0.22	17976	-2.08	-0.02
	18541	-1.02	-0.21	13194	-2.06	0.21
	17960	-0.19	-0.21	19557	-1.64	0.14
	18100	0.43	-0.21	13195	-1.52	-0.10
	17131	-0.34	-0.20	18543	-1.39	-0.63
	18099	0.46	-0.18	18466	-1.28	-0.14

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	18467	-2.16	-0.17	17983	-1.16	-0.15
	17133	0.84	-0.15	18463	-1.14	-0.58
	17983	-1.16	-0.15	19380	-1.14	0.02
	18466	-1.28	-0.14	19625	-1.09	0.23
	17958	-0.34	-0.12	19556	-1.03	-0.04
	17138	-0.94	-0.12	17984	-1.02	0.01
	13195	-1.52	-0.10	18541	-1.02	-0.21
	17986	-0.74	-0.10	18539	-0.96	0.15
	18538	-0.25	-0.08	19387	-0.95	-0.02
	17956	-0.74	-0.08	18450	-0.95	-0.35
	18465	-2.12	-0.07	17138	-0.94	-0.12
	17954	-2.20	-0.05	19381	-0.91	0.12
	19556	-1.03	-0.04	17135	-0.85	-0.25
	17976	-2.08	-0.02	19558	-0.82	0.04
	17139	-0.67	-0.02	19388	-0.79	0.51
	19387	-0.95	-0.02	19626	-0.76	0.13
	19632	-0.30	0.01	17956	-0.74	-0.08
	17984	-1.02	0.01	17986	-0.74	-0.10
	19380	-1.14	0.02	17139	-0.67	-0.02
	17955	-0.54	0.02	19379	-0.67	0.05
	19558	-0.82	0.04	19624	-0.62	0.18
	19389	-0.54	0.04	19634	-0.55	0.18
	19379	-0.67	0.05	19389	-0.54	0.04
	17975	-2.45	0.08	17955	-0.54	0.02
	17987	0.09	0.09	17132	-0.43	0.12
	17132	-0.43	0.12	17982	-0.38	-0.22
	19381	-0.91	0.12	17131	-0.34	-0.20
	19626	-0.76	0.13	17958	-0.34	-0.12
	19557	-1.64	0.14	19632	-0.30	0.01
	18539	-0.96	0.15	18538	-0.25	-0.08

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	19542	0.03	0.18	17960	-0.19	-0.21
	19624	-0.62	0.18	19541	-0.06	0.28
	19634	-0.55	0.18	18098	0.02	-0.29
	13194	-2.06	0.21	19542	0.03	0.18
	13193	-2.11	0.21	17987	0.09	0.09
	19633	0.31	0.22	19540	0.16	0.41
	19625	-1.09	0.23	17959	0.18	-0.38
	17974	-2.43	0.25	19633	0.31	0.22
	19541	-0.06	0.28	18100	0.43	-0.21
	18537	-2.88	0.29	18099	0.46	-0.18
	19540	0.16	0.41	17988	0.79	-0.48
	19388	-0.79	0.51	17133	0.84	-0.15

2.7.2.2 Grade 2-3

Table 2.7.2.2

Equating Summary: Read 2-3 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	72	0.08 (0.83)	72	0.06 (0.82)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-1.95	2.46	-1.81	2.46
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	60	0.11 (0.85)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	60	0.11 (0.85)		
	Percentage Anchors	**Average Displacement**		
	83%	-0.06		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	19400	-0.40	-0.70	19569	-1.81	0.28
	19392	-1.25	-0.59	17887	-1.41	0.03
	19393	-0.07	-0.58	17879	-1.41	-0.02
	19571	-0.67	-0.57	19399	-1.34	-0.43
	19646	0.14	-0.57	19392	-1.25	-0.59
	19645	-1.02	-0.52	19644	-1.04	-0.18
	18362	0.94	-0.46	19645	-1.02	-0.52
	19399	-1.34	-0.43	17888	-0.96	-0.05
	13346	0.60	-0.30	17886	-0.83	0.26
	13345	1.24	-0.29	19404	-0.82	0.12
	18475	0.16	-0.28	19401	-0.79	0.09
	18363	0.11	-0.26	18361	-0.67	-0.24
	19405	0.50	-0.24	19571	-0.67	-0.57
	18361	-0.67	-0.24	17880	-0.52	-0.04
	18473	0.01	-0.23	17153	-0.51	0.14

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	19570	0.76	-0.21	17950	-0.45	0.12
	19403	0.72	-0.19	17878	-0.44	0.39
	19644	-1.04	-0.18	19391	-0.41	-0.08
	17050	1.25	-0.17	19400	-0.40	-0.70
	13344	0.33	-0.15	19575	-0.26	0.21
	16095	0.70	-0.15	13340	-0.25	-0.06
	17049	1.22	-0.13	19574	-0.16	0.24
	16092	1.27	-0.13	18367	-0.10	0.21
	17894	0.23	-0.13	19393	-0.07	-0.58
	17892	0.42	-0.12	17154	-0.04	-0.08
	18474	0.25	-0.11	18473	0.01	-0.23
	13338	0.80	-0.10	17893	0.05	0.10
	17154	-0.04	-0.08	18363	0.11	-0.26
	19391	-0.41	-0.08	19646	0.14	-0.57
	13340	-0.25	-0.06	18475	0.16	-0.28
	13339	0.38	-0.06	17155	0.17	0.01
	17888	-0.96	-0.05	17894	0.23	-0.13
	17880	-0.52	-0.04	18474	0.25	-0.11
	17928	2.46	-0.04	17952	0.28	0.19
	17879	-1.41	-0.02	17051	0.32	0.06
	17155	0.17	0.01	13344	0.33	-0.15
	17887	-1.41	0.03	13339	0.38	-0.06
	16094	0.90	0.03	19573	0.39	0.26
	17051	0.32	0.06	17892	0.42	-0.12
	19401	-0.79	0.09	19405	0.50	-0.24
	17893	0.05	0.10	19654	0.55	0.35
	19404	-0.82	0.12	13346	0.60	-0.30
	17951	0.87	0.12	18366	0.64	0.19
	17950	-0.45	0.12	16095	0.70	-0.15
	17153	-0.51	0.14	19403	0.72	-0.19

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	17934	1.31	0.16	19570	0.76	-0.21
	18366	0.64	0.19	13338	0.80	-0.10
	17952	0.28	0.19	18365	0.83	0.38
	19652	0.98	0.20	17951	0.87	0.12
	18367	-0.10	0.21	16094	0.90	0.03
	19575	-0.26	0.21	18362	0.94	-0.46
	19653	1.26	0.22	19652	0.98	0.20
	19574	-0.16	0.24	17924	1.13	0.24
	17924	1.13	0.24	17049	1.22	-0.13
	17886	-0.83	0.26	13345	1.24	-0.29
	19573	0.39	0.26	17050	1.25	-0.17
	19569	-1.81	0.28	19653	1.26	0.22
	19654	0.55	0.35	16092	1.27	-0.13
	18365	0.83	0.38	17934	1.31	0.16
	17878	-0.44	0.39	17928	2.46	-0.04

2.7.2.3 Grades 4-5

Table 2.7.2.3

Equating Summary: Read 4-5 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	72	0.94 (1.08)	72	1.04 (1.10)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-2.04	2.99	-1.07	3.49
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	51	0.79 (0.96)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	51	0.79 (0.96)		
	Percentage Anchors	**Average Displacement**		
	71%	-0.03		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	19526	0.59	-0.85	18553	-1.07	0.01
	19524	0.69	-0.70	13407	-0.72	-0.02
	19763	2.25	-0.69	18409	-0.65	0.01
	19761	1.59	-0.30	18112	-0.59	0.18
	18198	0.71	-0.26	18119	-0.57	0.04
	17110	1.17	-0.25	18184	-0.52	-0.10
	15708	1.38	-0.23	18116	-0.37	0.17
	16010	1.34	-0.20	14626	-0.26	0.36
	19525	0.07	-0.19	13409	-0.17	0.07
	16011	0.12	-0.18	16009	-0.10	0.01
	15706	0.21	-0.16	18410	-0.06	0.10
	18411	0.80	-0.15	18554	-0.06	-0.10
	18555	0.17	-0.14	17109	-0.03	0.17
	16019	1.59	-0.12	13408	0.06	0.10
	18485	1.89	-0.12	19525	0.07	-0.19

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	19762	0.70	-0.12	16017	0.09	-0.11
	16017	0.09	-0.11	16011	0.12	-0.18
	18554	-0.06	-0.10	18555	0.17	-0.14
	18184	-0.52	-0.10	15706	0.21	-0.16
	18196	1.36	-0.08	14625	0.26	0.09
	18413	1.14	-0.05	19526	0.59	-0.85
	13407	-0.72	-0.02	18186	0.66	0.01
	18197	0.80	-0.01	19524	0.69	-0.70
	18186	0.66	0.01	19762	0.70	-0.12
	18553	-1.07	0.01	18198	0.71	-0.26
	18409	-0.65	0.01	18411	0.80	-0.15
	16018	1.28	0.01	18197	0.80	-0.01
	18487	2.15	0.01	18125	0.98	0.19
	16009	-0.10	0.01	18128	0.99	0.31
	18123	1.39	0.02	18413	1.14	-0.05
	18486	2.40	0.02	18185	1.15	0.22
	17111	1.47	0.03	17110	1.17	-0.25
	18119	-0.57	0.04	15707	1.23	0.04
	15707	1.23	0.04	16018	1.28	0.01
	18416	1.73	0.05	16010	1.34	-0.20
	13409	-0.17	0.07	18196	1.36	-0.08
	14625	0.26	0.09	15708	1.38	-0.23
	18410	-0.06	0.10	18123	1.39	0.02
	13408	0.06	0.10	17111	1.47	0.03
	18415	2.44	0.12	19761	1.59	-0.30
	18116	-0.37	0.17	14627	1.59	0.32
	17109	-0.03	0.17	16019	1.59	-0.12
	18112	-0.59	0.18	19757	1.68	0.25
	18125	0.98	0.19	18416	1.73	0.05
	18185	1.15	0.22	18485	1.89	-0.12

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	19758	2.54	0.25	18487	2.15	0.01
	19757	1.68	0.25	19763	2.25	-0.69
	18128	0.99	0.31	18486	2.40	0.02
	14627	1.59	0.32	18415	2.44	0.12
	19759	2.68	0.34	19758	2.54	0.25
	14626	-0.26	0.36	19759	2.68	0.34

2.7.2.4 Grades 6-8

Table 2.7.2.4

Equating Summary: Read 6-8 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	72	1.48 (1.31)	72	1.47 (1.32)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-1.69	3.79	-1.69	4.05
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	51	1.07 (1.22)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	51	1.07 (1.22)		
	Percentage Anchors	**Average Displacement**		
	71%	0.00		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	18505	1.33	-0.46	18062	-1.69	0.04
	19686	1.26	-0.45	13575	-1.58	0.66
	18420	1.04	-0.43	19684	-1.09	-0.03
	18383	1.58	-0.41	18417	-0.84	0.12
	18382	0.08	-0.40	19616	-0.77	0.12
	19618	0.83	-0.39	19472	-0.67	0.06
	18328	2.50	-0.31	19617	-0.60	0.02
	18322	1.88	-0.31	13576	-0.48	0.61
	18507	2.47	-0.25	18321	-0.28	0.17
	19685	-0.13	-0.15	19685	-0.13	-0.15
	13962	1.39	-0.14	13577	-0.09	0.38
	13615	2.21	-0.14	18382	0.08	-0.40
	18506	1.58	-0.13	18381	0.22	0.03
	19474	0.75	-0.13	19473	0.38	-0.10
	19473	0.38	-0.10	18063	0.44	0.24

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	13614	1.95	-0.06	18064	0.57	0.11
	18319	1.88	-0.06	18424	0.74	-0.01
	18055	1.14	-0.05	19474	0.75	-0.13
	13616	2.67	-0.04	13629	0.78	-0.04
	13629	0.78	-0.04	19618	0.83	-0.39
	13963	1.94	-0.04	13631	0.84	0.27
	19684	-1.09	-0.03	19485	0.86	0.05
	19484	2.37	-0.01	18420	1.04	-0.43
	18424	0.74	-0.01	13650	1.09	0.15
	18056	1.17	0.01	18055	1.14	-0.05
	18327	2.26	0.01	18056	1.17	0.01
	19617	-0.60	0.02	19686	1.26	-0.45
	18381	0.22	0.03	18505	1.33	-0.46
	18062	-1.69	0.04	18054	1.38	0.11
	19485	0.86	0.05	13962	1.39	-0.14
	19700	2.13	0.06	13630	1.47	0.20
	19472	-0.67	0.06	18383	1.58	-0.41
	13651	2.53	0.07	18506	1.58	-0.13
	18054	1.38	0.11	18319	1.88	-0.06
	18064	0.57	0.11	18322	1.88	-0.31
	18417	-0.84	0.12	13963	1.94	-0.04
	19616	-0.77	0.12	13614	1.95	-0.06
	13652	2.41	0.13	18318	2.01	0.15
	13964	2.02	0.15	13964	2.02	0.15
	13650	1.09	0.15	19700	2.13	0.06
	18318	2.01	0.15	13615	2.21	-0.14
	18321	-0.28	0.17	18327	2.26	0.01
	13630	1.47	0.20	19484	2.37	-0.01
	19486	2.98	0.21	13652	2.41	0.13
	19701	2.58	0.21	18507	2.47	-0.25

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	18063	0.44	0.24	18328	2.50	-0.31
	13631	0.84	0.27	13651	2.53	0.07
	19702	3.15	0.31	19701	2.58	0.21
	13577	-0.09	0.38	13616	2.67	-0.04
	13576	-0.48	0.61	19486	2.98	0.21
	13575	-1.58	0.66	19702	3.15	0.31

2.7.2.5 Grades 9-12

Table 2.7.2.5

Equating Summary: Read 9-12 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Items	**Average Difficulty (Std. Dev.)**	**No. of Items**	**Average Difficulty (Std. Dev.)**
	72	2.34 (1.31)	72	2.28 (1.19)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-1.20	4.52	-1.20	4.52
Anchoring Items	**No. of Possible Anchors**	**Average Difficulty (Std. Dev.)**		
	51	2.18 (1.28)		
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**		
	51	2.18 (1.28)		
	Percentage Anchors	**Average Displacement**		
	71%	-0.01		

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	**Item Difficulty**	**Displacement**	**Item ID**	**Item Difficulty**	**Displacement**
	18447	2.11	-0.79	18431	-1.20	0.37
	18457	3.00	-0.54	18432	0.10	0.06
	19662	2.90	-0.47	18509	0.26	0.22
	18511	1.86	-0.37	18445	0.44	0.35
	18456	2.82	-0.37	17998	0.45	0.02
	19661	0.60	-0.33	17996	0.59	0.08
	18510	0.60	-0.32	18510	0.60	-0.32
	17075	3.98	-0.24	19661	0.60	-0.33
	13950	1.41	-0.17	18433	0.90	0.00
	17077	4.52	-0.16	19453	0.92	-0.07
	18519	3.06	-0.14	17999	0.99	0.19
	19452	1.25	-0.11	19660	1.09	-0.07
	16072	3.26	-0.10	18446	1.10	0.11
	18518	2.79	-0.07	18023	1.11	0.22
	19660	1.09	-0.07	19452	1.25	-0.11

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	19453	0.92	-0.07	19454	1.27	-0.04
	19454	1.27	-0.04	13950	1.41	-0.17
	16070	2.65	-0.04	18455	1.53	0.02
	19672	3.11	-0.03	18517	1.58	0.11
	19466	3.96	-0.03	18511	1.86	-0.37
	19464	2.96	-0.02	18025	1.88	0.11
	18526	3.40	-0.01	13952	1.92	0.16
	18433	0.90	0.00	18447	2.11	-0.79
	19673	3.14	0.01	18024	2.25	0.18
	18455	1.53	0.02	18030	2.25	0.21
	17998	0.45	0.02	16071	2.36	0.12
	18527	2.95	0.04	17076	2.40	0.14
	13951	2.69	0.05	19465	2.41	0.27
	18432	0.10	0.06	16070	2.65	-0.04
	19596	3.25	0.06	13951	2.69	0.05
	17996	0.59	0.08	18518	2.79	-0.07
	18032	3.45	0.08	18456	2.82	-0.37
	18025	1.88	0.11	19662	2.90	-0.47
	18446	1.10	0.11	18527	2.95	0.04
	18517	1.58	0.11	19464	2.96	-0.02
	16071	2.36	0.12	18457	3.00	-0.54
	19597	3.56	0.13	18519	3.06	-0.14
	18525	4.09	0.13	19672	3.11	-0.03
	17076	2.40	0.14	19673	3.14	0.01
	13952	1.92	0.16	19596	3.25	0.06
	18031	3.57	0.16	16072	3.26	-0.10
	18024	2.25	0.18	18526	3.40	-0.01
	17999	0.99	0.19	18032	3.45	0.08
	18030	2.25	0.21	19597	3.56	0.13
	18023	1.11	0.22	18031	3.57	0.16

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	18509	0.26	0.22	19598	3.82	0.25
	19674	3.97	0.25	19466	3.96	-0.03
	19598	3.82	0.25	19674	3.97	0.25
	19465	2.41	0.27	17075	3.98	-0.24
	18445	0.44	0.35	18525	4.09	0.13
	18431	-1.20	0.37	17077	4.52	-0.16

2.7.3 Writing

2.7.3.1 Grade 1

Table 2.7.3.1.1

Equating Summary: Writ 1 A S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	-0.41 (0.24)	2	-0.38 (0.29)
	Easiest	Hardest	Easiest	Hardest
	-0.58	-0.24	-0.58	-0.18
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)		
	1	-0.58 (N/A)		
	No. of Anchors Used	Average Difficulty (Std. Dev.)		
	1	-0.58 (N/A)		
	Percentage Anchors*	Average Displacement		
	50%	0.00		
Rating Scale Step Measures by Task	Anchored Scale Steps			
	Step	Measure		
	1	-2.47		
	2	-2.78		
	3	-2.61		
	4	-1.68		
	5	-0.48		
	6	0.97		
	7	2.25		
	8	3.21		
	9	3.59		
Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19414	-0.58	0.00	19414	-0.58	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

Table 2.7.3.1.2

Equating Summary: Writ 1 B/C S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	-0.20 (0.16)	2	-0.08 (0.33)
	Easiest	Hardest	Easiest	Hardest
	-0.31	-0.09	-0.31	0.16

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	-0.31 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	-0.31 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19213	-0.31	0.00	19213	-0.31	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

2.7.3.2 Grade 2-3

Table 2.7.3.2.1

Equating Summary: Writ 2-3 A S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	-0.05 (0.05)	2	-0.01 (0.11)
	Easiest	Hardest	Easiest	Hardest
	-0.09	-0.02	-0.09	0.07

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	-0.09 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	-0.09 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	18984	-0.09	0.00	18984	-0.09	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

Table 2.7.3.2.2

Equating Summary: Writ 2-3 B/C S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	0.28 (0.06)	2	0.28 (0.06)
	Easiest	Hardest	Easiest	Hardest
	0.25	0.32	0.23	0.32

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	0.32 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	0.32 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19178	0.32	0.00	19178	0.32	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

2.7.3.3 Grades 4-5

Table 2.7.3.3.1

Equating Summary: Writ 4-5 A S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	1.13 (0.17)	2	1.11 (0.14)
	Easiest	Hardest	Easiest	Hardest
	1.01	1.24	1.01	1.21

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	1.01 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	1.01 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	18976_19735	1.01	0.00	18976_19735	1.01	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating,
(number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

Table 2.7.3.3.2

Equating Summary: Writ 4-5 B/C S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	2.25 (0.38)	2	2.50 (0.02)
	Easiest	Hardest	Easiest	Hardest
	1.98	2.52	2.49	2.52

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	2.52 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	2.52 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19222_19737	2.52	0.00	19222_19737	2.52	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

2.7.3.4 Grades 6-8

Table 2.7.3.4.1

Equating Summary: Writ 6-8 A S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	0.76 (0.22)	2	0.75 (0.21)
	Easiest	Hardest	Easiest	Hardest
	0.60	0.91	0.60	0.90

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	0.60 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	0.60 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	18969_19742	0.60	0.00	18969_19742	0.60	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

Table 2.7.3.4.2

Equating Summary: Writ 6-8 B/C S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	1.26 (0.27)	2	1.29 (0.32)
	Easiest	Hardest	Easiest	Hardest
	1.07	1.45	1.07	1.52

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	1.07 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	1.07 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19212_19740	1.07	0.00	19212_19740	1.07	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

2.7.3.5 Grades 9-12

Table 2.7.3.5.1

Equating Summary: Writ 9-12 A S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	2.12 (0.07)	2	2.18 (0.17)
	Easiest	Hardest	Easiest	Hardest
	2.06	2.17	2.06	2.31
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)		
	1	2.06 (N/A)		
	No. of Anchors Used	Average Difficulty (Std. Dev.)		
	1	2.06 (N/A)		
	Percentage Anchors*	Average Displacement		
	50%	0.00		
Common Rating Scale Step Measures	Anchored Scale Steps			
	Step	Measure		
	1	-2.47		
	2	-2.78		
	3	-2.61		
	4	-1.68		
	5	-0.48		
	6	0.97		
	7	2.25		
	8	3.21		
	9	3.59		

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	18989_19745	2.06	0.00	18989_19745	2.06	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating, (number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

Table 2.7.3.5.2

Equating Summary: Writ 9-12 B/C S601 Online

Comparison of Forms*	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	2	2.00 (0.45)	2	2.02 (0.42)
	Easiest	Hardest	Easiest	Hardest
	1.68	2.32	2.00	2.02

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	1	2.32 (N/A)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	1	2.32 (N/A)
	Percentage Anchors*	Average Displacement
	50%	0.00

Common Rating Scale Step Measures	Anchored Scale Steps	
	Step	Measure
	1	-2.47
	2	-2.78
	3	-2.61
	4	-1.68
	5	-0.48
	6	0.97
	7	2.25
	8	3.21
	9	3.59

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	17319_18252	2.32	0.00	17319_18252	2.32	0.00

* includes a new task and one anchor tasks (i.e., one continuing) used for form 601 equating,
(number of anchor tasks)/(number of new tasks + number of anchor tasks) = 1/2.

2.7.4 Speaking

2.7.4.1 Grade 1

Table 2.7.4.1

Equating Summary: Spek 1 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	9	-1.71 (2.27)	12*	-1.65 (2.07)
	Easiest	Hardest	Easiest	Hardest
	-4.76	0.61	-4.62	0.31

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	3	-1.69 (2.54)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	3	-1.69 (2.54)
	Percentage Anchors	Average Displacement
	33%	0.00

Rating Scale Step Measures by Task	Anchored Scale Steps		
	Task	Step	Measure
	PL 1 Tasks	1	0.56
		2	-0.56
	PL 3/PL 5 Tasks	1	-2.65
		2	-1.80
		3	1.46
		4	2.98

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19129	-0.33	-0.04	19124	-4.62	-0.02
	19124	-4.62	-0.02	19129	-0.33	-0.04
	19136	-0.11	0.07	19136	-0.11	0.07

* Note: operational and external anchor tasks included in total number of tasks.

See S503 Online ATR, Section 2.7, for more information.

2.7.4.2 Grade 2-3

Table 2.7.4.2

Equating Summary: Spek 2-3 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	**Average Difficulty (Std. Dev.)**	**No. of Tasks**	**Average Difficulty (Std. Dev.)**
	9	-1.43 (2.53)	12*	-1.63 (2.39)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-4.95	0.66	-5.00	0.56

Anchoring Tasks		
	No. of Possible Anchors	**Average Difficulty (Std. Dev.)**
	3	-1.59 (2.95)
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**
	3	-1.59 (2.95)
	Percentage Anchors	**Average Displacement**
	33%	0.00

Rating Scale Step Measures by Task	Anchored Scale Steps		
	Task	**Step**	**Measure**
	PL 1 Tasks	1	0.56
		2	-0.56
	PL 3/PL 5 Tasks	1	-2.65
		2	-1.80
		3	1.46
		4	2.98

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	**Task Difficulty**	**Displacement**	**Task ID**	**Task Difficulty**	**Displacement**
	19149	-0.39	-0.12	19144	-4.95	-0.02
	19144	-4.95	-0.02	19149	-0.39	-0.12
	19164	0.56	0.15	19164	0.56	0.15

* Note: operational and external anchor tasks included in total number of tasks.

See S503 Online ATR, Section 2.7, for more information.

2.7.4.3 Grades 4-5

Table 2.7.4.3

Equating Summary: Spek 4-5 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	**Average Difficulty (Std. Dev.)**	**No. of Tasks**	**Average Difficulty (Std. Dev.)**
	9	-0.22 (2.83)	12*	-0.31 (2.69)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-4.02	2.35	-4.23	2.06

Anchoring Tasks		
	No. of Possible Anchors	**Average Difficulty (Std. Dev.)**
	3	-0.40 (3.16)
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**
	3	-0.40 (3.16)
	Percentage Anchors	**Average Displacement**
	33%	-0.03

Rating Scale Step Measures by Task	Anchored Scale Steps		
	Task	**Step**	**Measure**
	PL 1 Tasks	1	0.56
		2	-0.56
	PL 3/PL 5 Tasks	1	-2.65
		2	-1.80
		3	1.46
		4	2.98

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	**Task Difficulty**	**Displacement**	**Task ID**	**Task Difficulty**	**Displacement**
	19076	-4.02	-0.10	19076	-4.02	-0.10
	19081	1.01	0.00	19081	1.01	0.00
	19088	1.81	0.01	19088	1.81	0.01

* Note: operational and external anchor tasks included in total number of tasks.

See S503 Online ATR, Section 2.7, for more information.

2.7.4.4 Grades 6-8

Table 2.7.4.4

Equating Summary: Spek 6-8 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)
	9	0.20 (2.61)	12*	0.30 (2.69)
	Easiest	Hardest	Easiest	Hardest
	-3.42	2.42	-3.52	2.97

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	5	0.94 (2.46)
	No. of Anchors Used	Average Difficulty (Std. Dev.)
	5	0.94 (2.46)
	Percentage Anchors	Average Displacement
	56%	-0.07

Rating Scale Step Measures by Task	Anchored Scale Steps		
	Task	Step	Measure
	PL 1 Tasks	1	0.56
		2	-0.56
	PL 3/PL 5 Tasks	1	-2.65
		2	-1.80
		3	1.46
		4	2.98

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	Task Difficulty	Displacement	Task ID	Task Difficulty	Displacement
	19139	-3.42	-0.37	19139	-3.42	-0.37
	19157	1.85	-0.23	19928	1.63	0.28
	19166	2.23	-0.19	19157	1.85	-0.23
	19935	2.42	0.15	19166	2.23	-0.19
	19928	1.63	0.28	19935	2.42	0.15

* Note: operational and external anchor tasks included in total number of tasks.

See S503 Online ATR, Section 2.7, for more information.

2.7.4.5 Grades 9-12

Table 2.7.4.5

Equating Summary: Spek 9-12 S601 Online

Comparison of Forms	Form 601		Form 503	
	No. of Tasks	**Average Difficulty (Std. Dev.)**	**No. of Tasks**	**Average Difficulty (Std. Dev.)**
	9	0.37 (2.51)	12*	0.58 (2.45)
	Easiest	**Hardest**	**Easiest**	**Hardest**
	-3.08	2.88	-2.85	2.88

Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)
	4	0.82 (2.41)
	No. of Anchors Used	**Average Difficulty (Std. Dev.)**
	4	0.82 (2.41)
	Percentage Anchors	**Average Displacement**
	44%	-0.07

Rating Scale Step Measures by Task	Anchored Scale Steps		
	Task	**Step**	**Measure**
	PL 1 Tasks	1	0.56
		2	-0.56
	PL 3/PL 5 Tasks	1	-2.65
		2	-1.80
		3	1.46
		4	2.98

Displacement of Anchor Tasks	Anchor Tasks by Displacement			Anchor Tasks by Task Difficulty		
	Task ID	**Task Difficulty**	**Displacement**	**Task ID**	**Task Difficulty**	**Displacement**
	19106	-2.66	-0.27	19106	-2.66	-0.27
	19189	1.76	-0.21	19013	1.30	0.32
	19194	2.88	-0.14	19189	1.76	-0.21
	19013	1.30	0.32	19194	2.88	-0.14

* Note: operational and external anchor tasks included in total number of tasks.

See S503 Online ATR, Section 2.7, for more information.

2.8 Test Characteristic Curve

Test characteristic curves (TCC) graphically show the functional relationship between a student's ability measure (in logits) on the horizontal axis and that student's expected raw score (i.e., the estimated true score) on the vertical axis. Thus, for a given ability measure, the corresponding expected raw score can be found via the TCC. For reporting purposes, WIDA uses the student's ability measure to determine the proficiency level. Since the TCC transforms ability measures to expected raw scores, this representation allows test users to relate a student's ability measure to his/her proficiency level (i.e., a more familiar frame of reference that test users employ to interpret students' scores), based on that student's expected total raw score.

Mathematically, the TCC is the sum of all item/task characteristic functions for the items and tasks included on the test form (Lord, 1980). Thus, the TCC depends on the item/task characteristic functions (Lord, 1980). The shape of the TCC depends on several factors, including the number and the characteristics of the items/tasks, the item response theory model used, and the values of the item/task parameters. Consequently, there is no explicit formula for the TCC, and there are no parameters for the curve (Baker & Kim, 2017). As we present the Listening and Reading Online ACCESS tests in a multistage adaptive format and they are not fixed test forms, it is not appropriate to present TCCs for these tests.

Since raters use a polytomous scoring scale for Writing and Speaking tasks, the shapes of the TCCs for these tests are also affected by the parameter values for the individual categories on the scoring tools that raters use to evaluate students' responses to the tasks. These scoring tools have more score categories than the scoring schemes used for evaluating students' responses to multiple-choice items, which we typically score using just two categories— "right" or "wrong." By contrast, the Writing and Speaking rating scales have multiple score categories. For Writing, the rating scale has six whole score categories with an additional three in-between "plus" score categories, for a total of nine possible score points; for Speaking, the rating scale has five score categories. Therefore, the student ability measures for the Writing and Speaking domains will span a wide logit range (e.g., for the Grade 1 Writing test, the student ability measures shown on the horizontal axis of Figure 2.8.3.1.1 range from -7 logits to 8 logits, a 15-logit spread).

Ideally, a TCC will be a smooth monotonically, or continuously increasing, S-shaped probability curve. However, when raters use multicategory rating scales to evaluate students' responses, they frequently do not assign equal numbers of scores in each of those categories. Consequently, the resulting adjacent score category boundaries may not be equidistant, and, indeed, in some cases, they may even be far apart if raters assign few scores in certain categories. In this situation, the curve of the TCC is likely to be somewhat bumpy or uneven across the student ability continuum. (The closer the adjacent score category boundaries are, the smoother the rise of the TCC along the student ability continuum.) Additionally, for some tests, the TCC may rise in a smooth S-shaped curve over the initial segment of the student ability continuum, but then plateau in the area between the boundaries of adjacent score categories before rising smoothly again, which would reflect the raters' uneven use of the score categories on the rating scale. We see this pattern in the TCCs for the Writing and Speaking tests. The TCCs for other tests that include open-ended tasks, such as the National Assessment of Educational Progress Writing assessment (Muraki, 1993), often have this shape.

There are five vertical lines in each of the TCC figures indicating, for each test form, the cut scores for the highest grade in each grade-level cluster, dividing each figure into six sections that denote the WIDA proficiency levels (PLs 1–6) for the domain. As would be expected, higher raw scores are required for placement in higher proficiency levels. The relative width of each section between the cut score lines gives an indication of how many raw score points a student must achieve to be placed into a WIDA proficiency level.

2.8.1 Listening

The ACCESS 2.0 Online Listening test is a multistage adaptive assessment. As students do not all take the same set of items in the test, no test characteristic curve is presented.

2.8.2 Reading

The ACCESS 2.0 Online Reading test is a multistage adaptive assessment. As students do not all take the same set of items in the test, no test characteristic curve is presented.

2.8.3 Writing

2.8.3.1 Grade 1

Figure 2.8.3.1.1 Test Characteristic Curve: Writ 1 A S601 Online

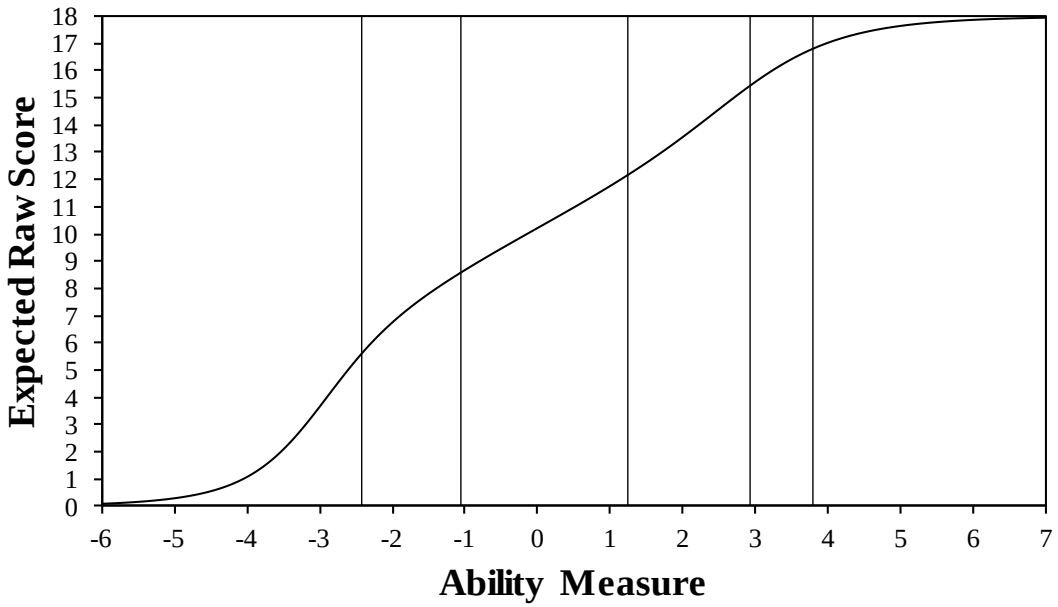
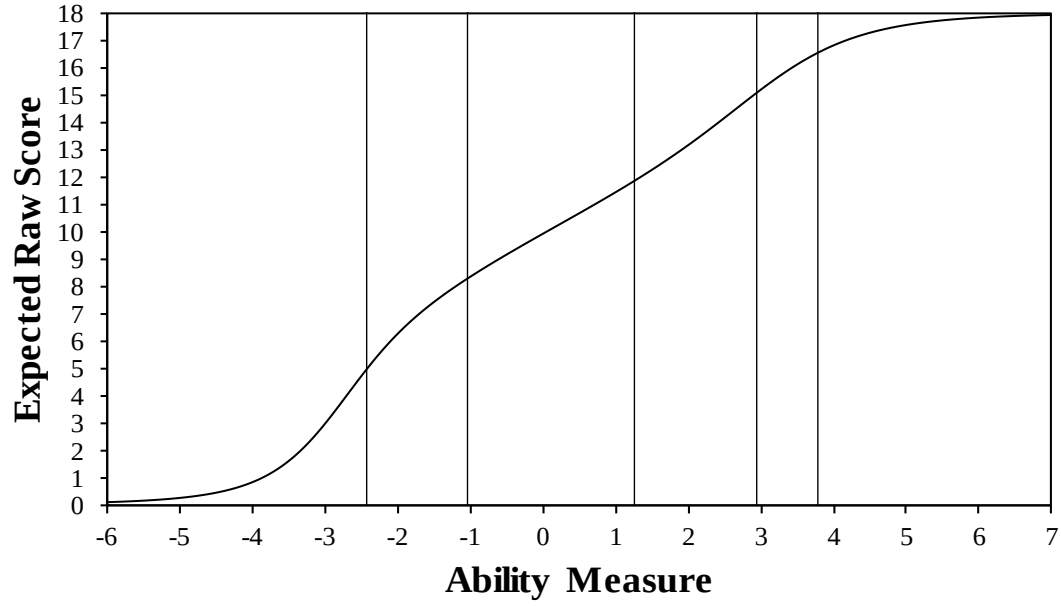


Figure 2.8.3.1.2 Test Characteristic Curve: Writ 1 B/C S601 Online



2.8.3.2 Grade 2-3

Figure 2.8.3.2.1 Test Characteristic Curve: Writ 2-3 A S601 Online

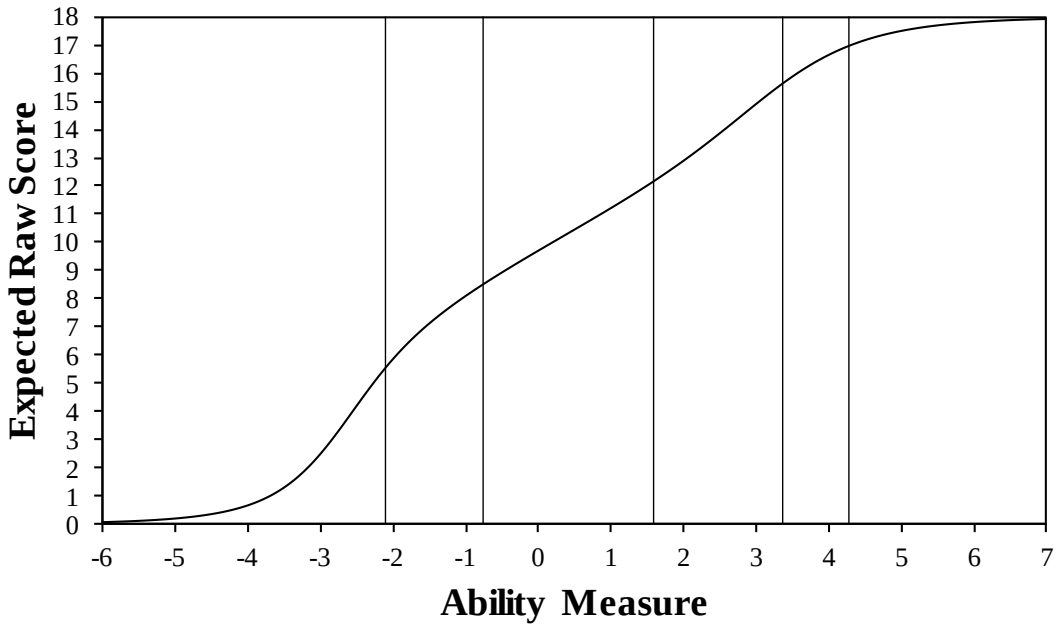
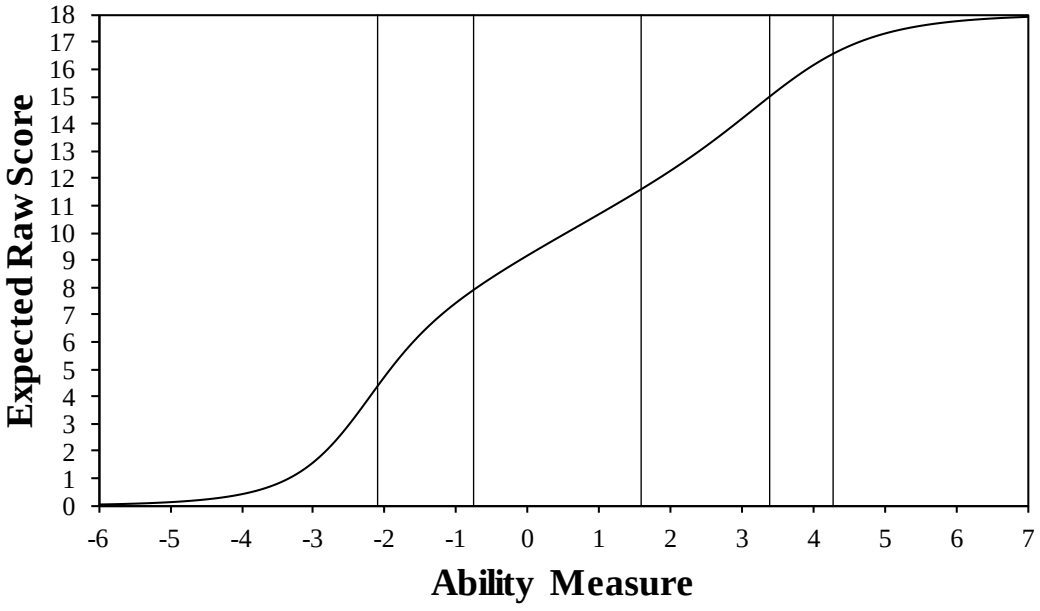


Figure 2.8.3.2.2 Test Characteristic Curve: Writ 2-3 B/C S601 Online



2.8.3.3 Grades 4-5

Figure 2.8.3.3.1 Test Characteristic Curve: Writ 4-5 A S601 Online

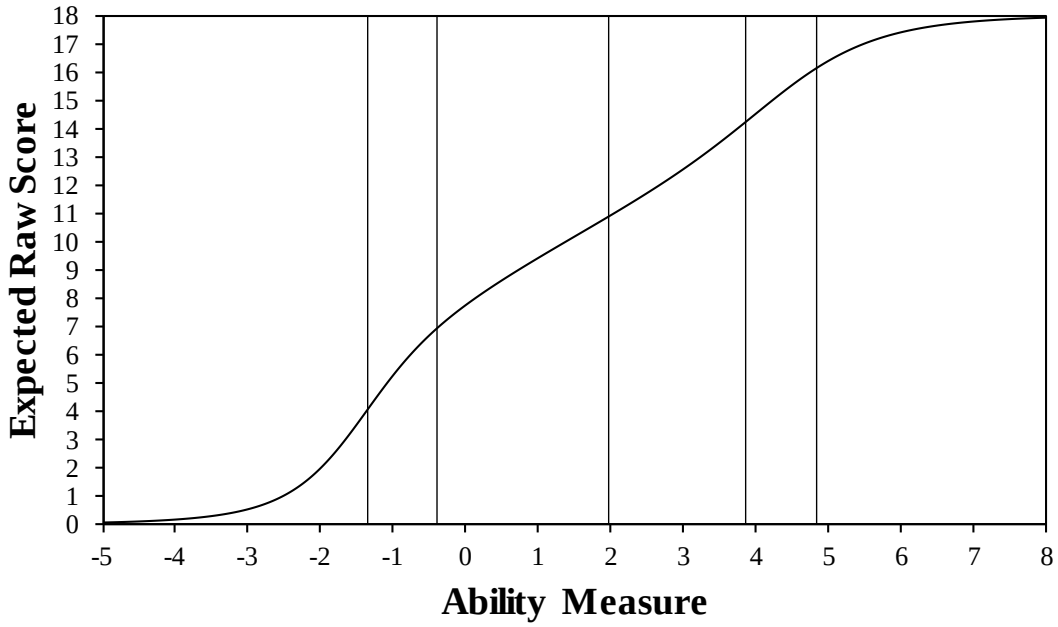
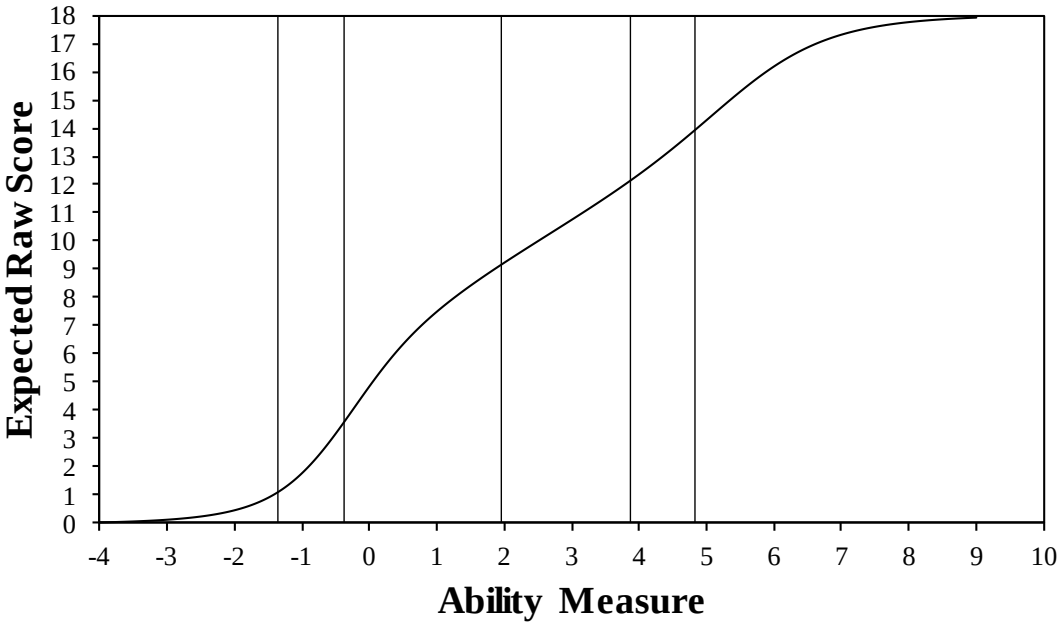


Figure 2.8.3.3.2 Test Characteristic Curve: Writ 4-5 B/C S601 Online



2.8.3.4 Grades 6-8

Figure 2.8.3.4.1 Test Characteristic Curve: Writ 6-8 A S601 Online

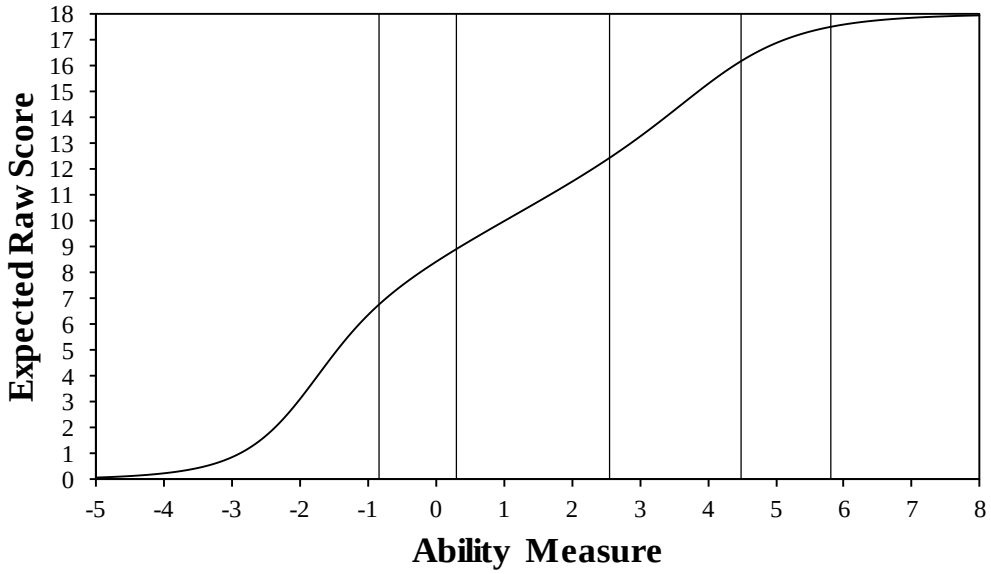
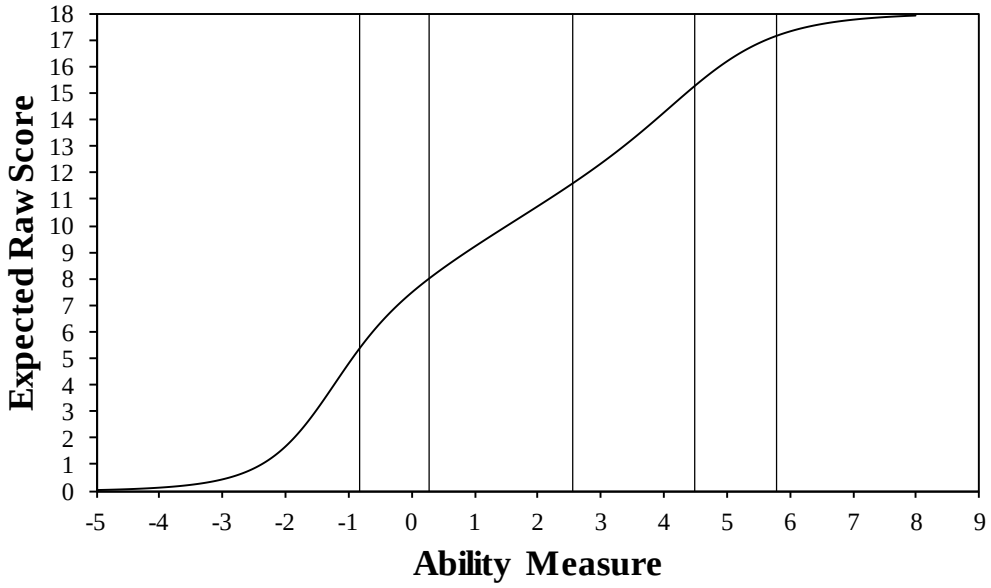


Figure 2.8.3.4.2 Test Characteristic Curve: Writ 6-8 B/C S601 Online



2.8.3.5 Grades 9-12

Figure 2.8.3.5.1 Test Characteristic Curve: Writ 9-12 A S601 Online

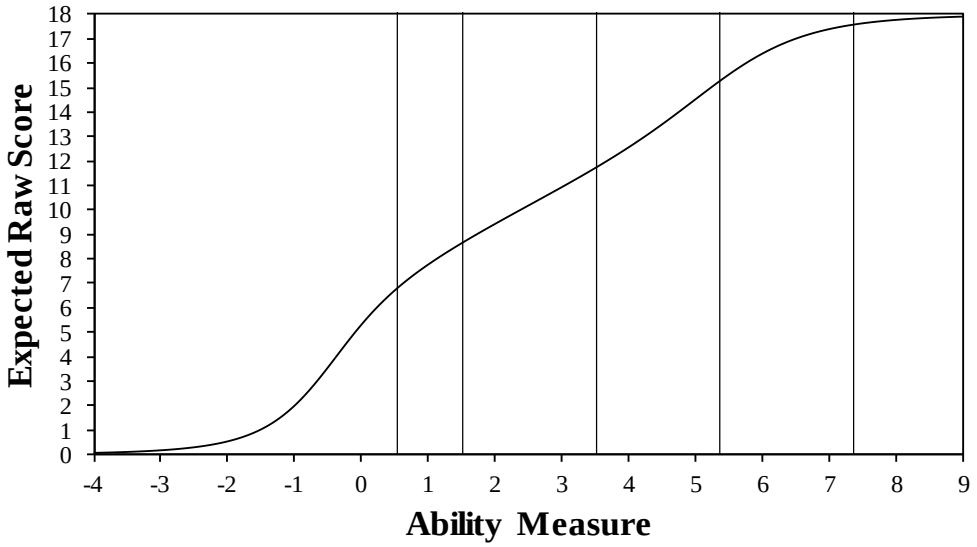
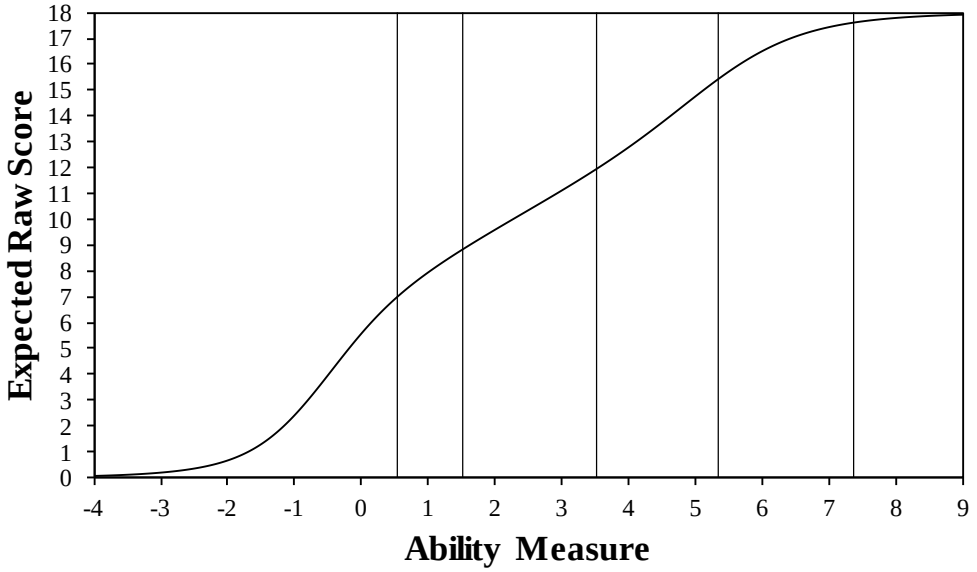


Figure 2.8.3.5.2 Test Characteristic Curve: Writ 9-12 B/C S601 Online



2.8.4 Speaking

2.8.4.1 Grade 1

Figure 2.8.4.1.1
Test Characteristic Curve: Spek 1 Pre-A S601 Online

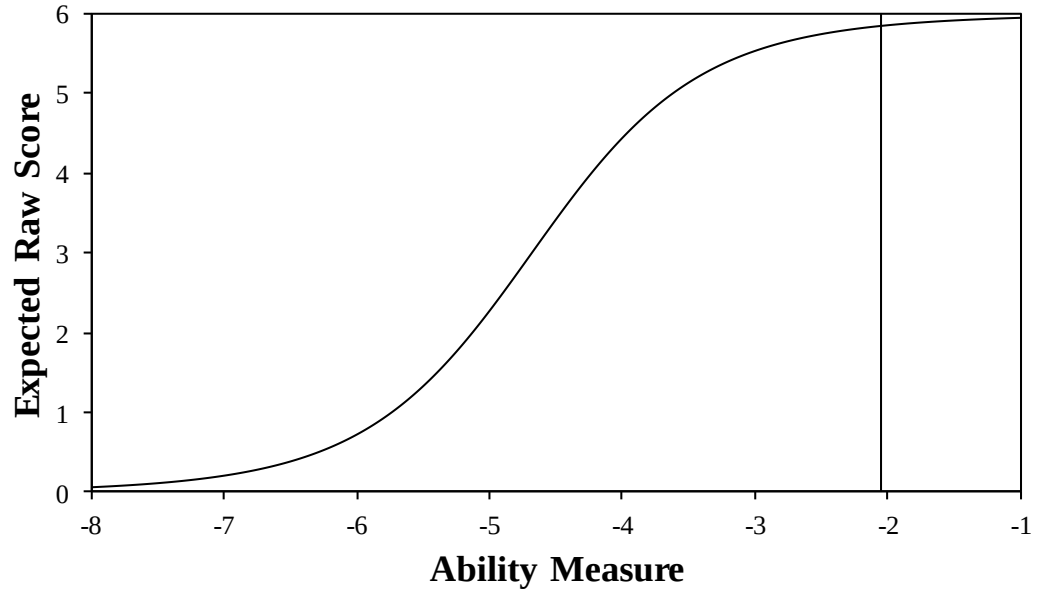


Figure 2.8.4.1.2
Test Characteristic Curve: Spek 1 A S601 Online

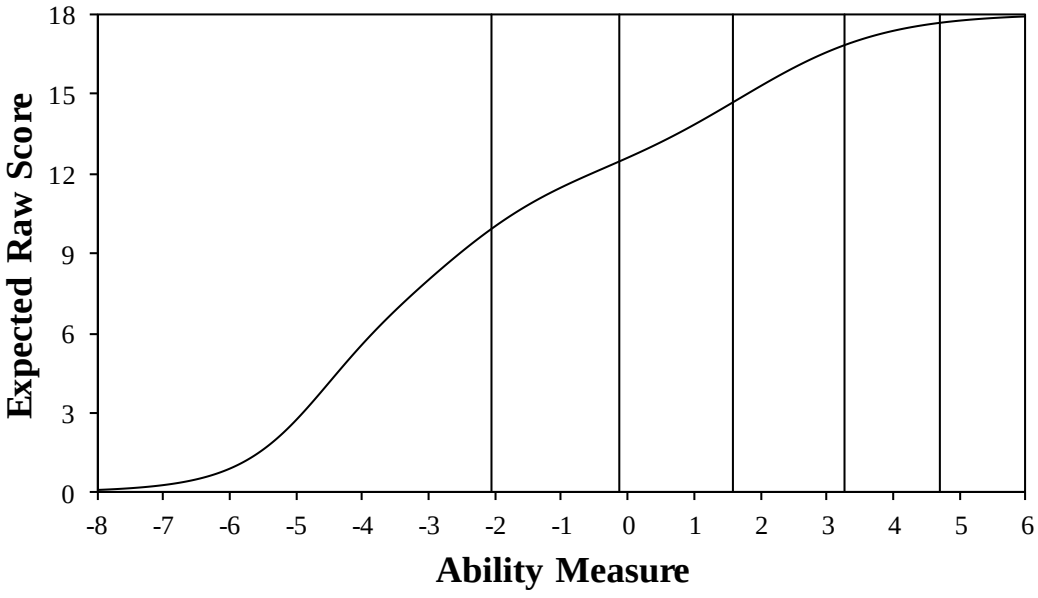
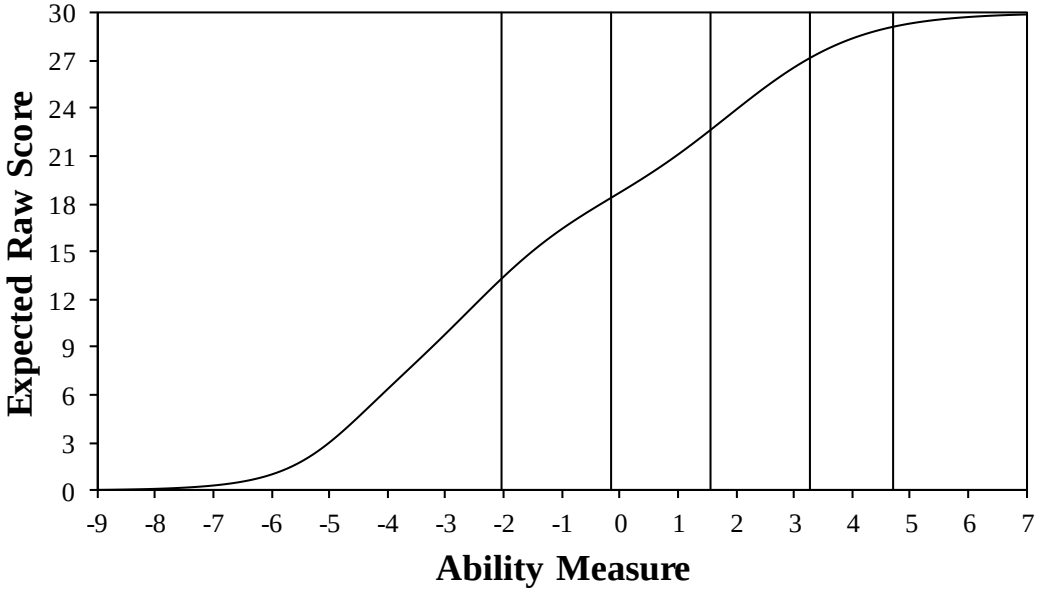


Figure 2.8.4.1.3
Test Characteristic Curve: Spek 1 B/C S601 Online



2.8.4.2 Grades 2-3

Figure 2.8.4.2.1
Test Characteristic Curve: Spek 2-3 Pre-A S601 Online

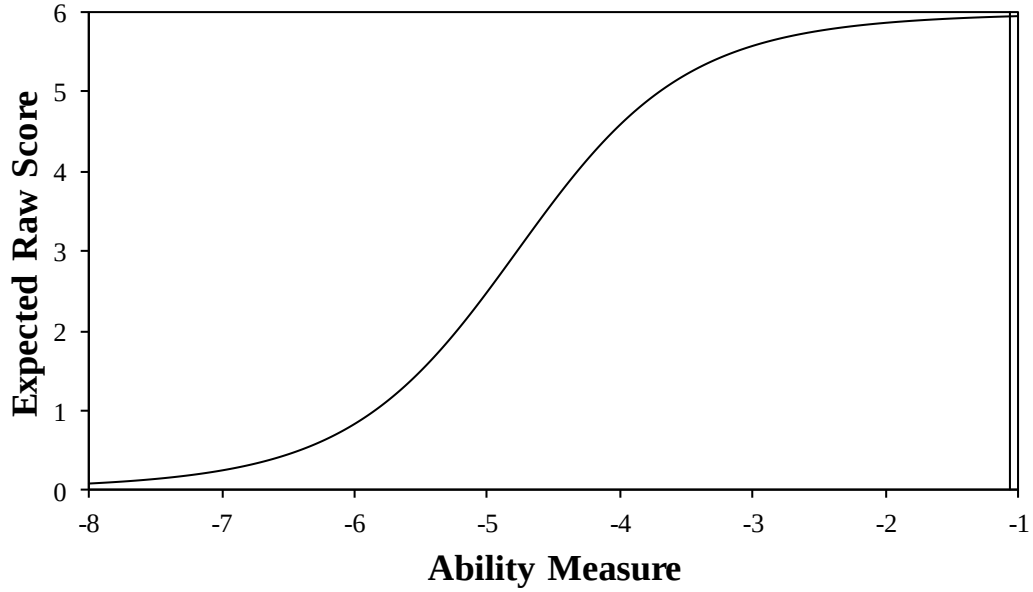


Figure 2.8.4.2.2
Test Characteristic Curve: Spek 2-3 A S601 Online

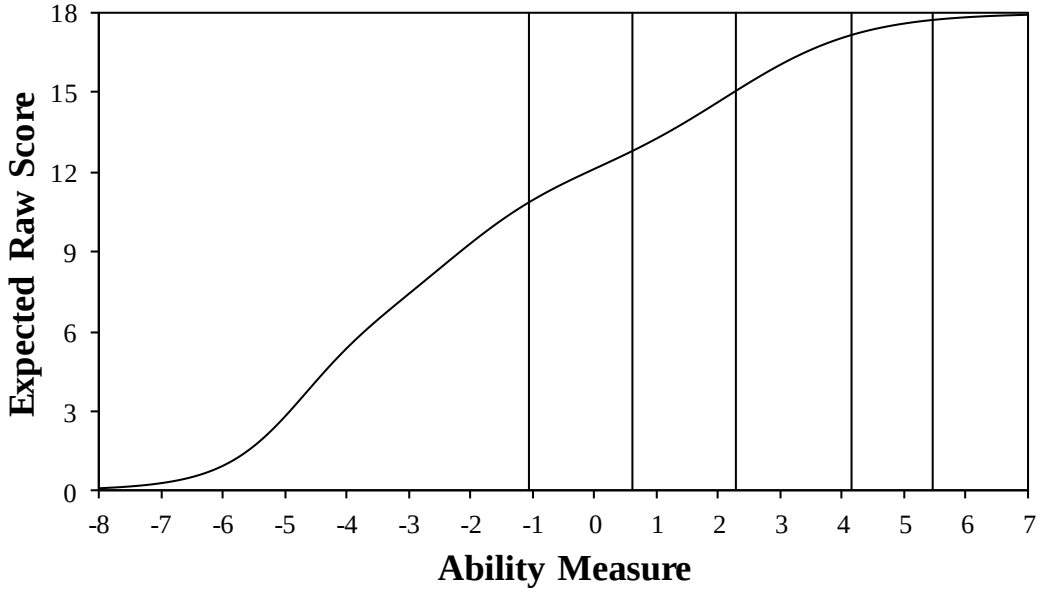
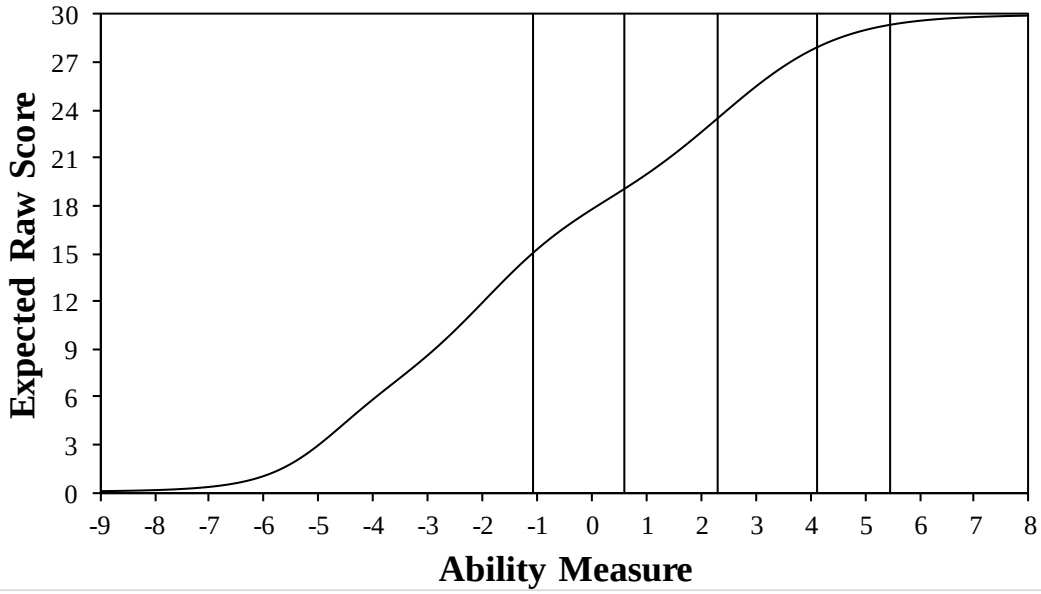


Figure 2.8.4.2.3
Test Characteristic Curve: Spek 2-3 B/C S601 Online



2.8.4.3 Grades 4-5

Figure 2.8.4.3.1
Test Characteristic Curve: Spek 4-5 Pre-A S601 Online

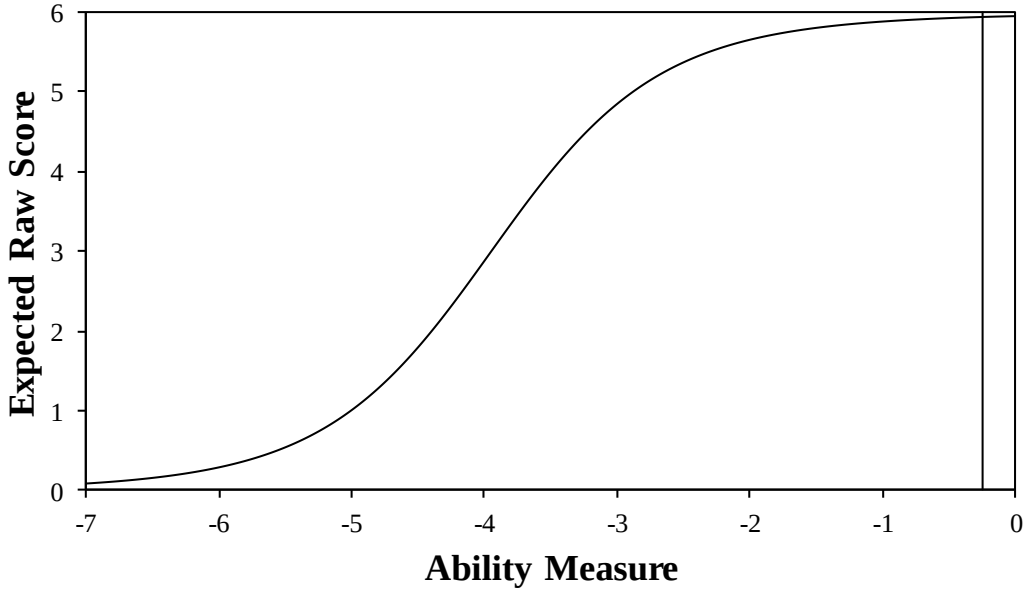


Figure 2.8.4.3.2
Test Characteristic Curve: Spek 4-5 A S601 Online

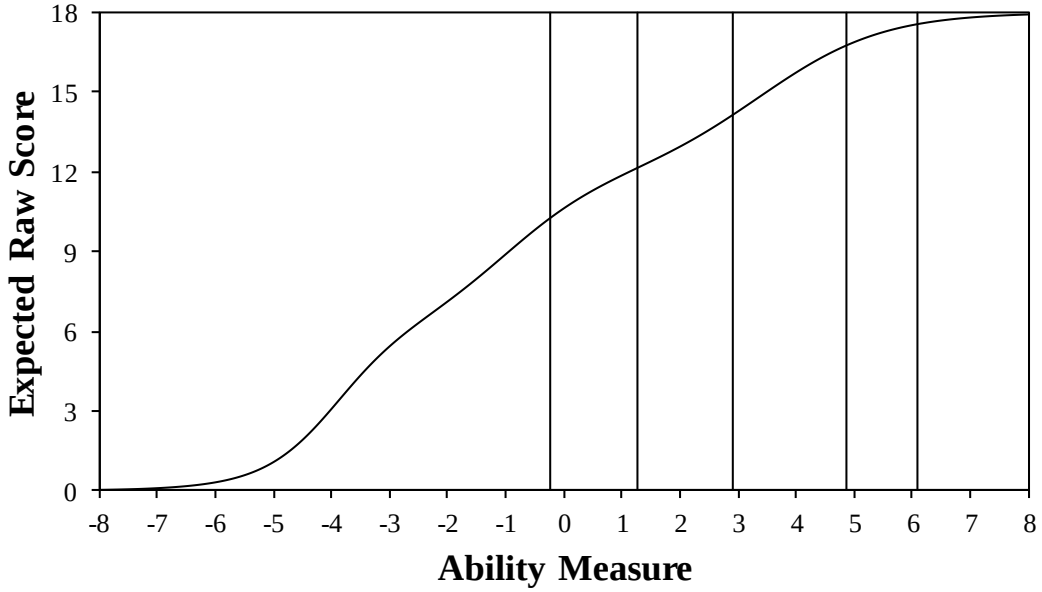
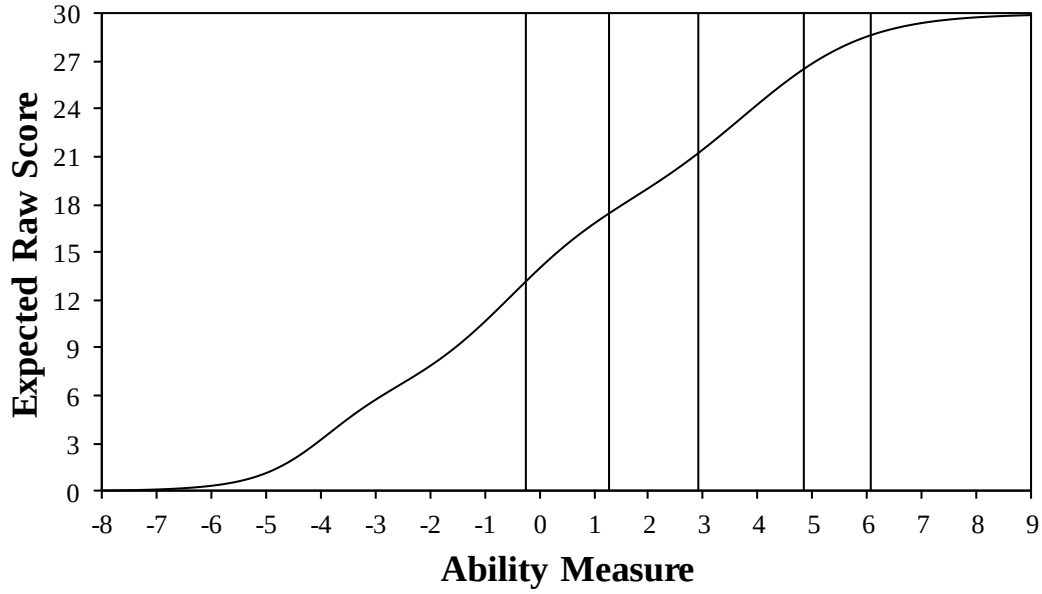


Figure 2.8.4.3.3
Test Characteristic Curve: Spek 4-5 B/C S601 Online



2.8.4.4 Grades 6-8

Figure 2.8.4.4.1
Test Characteristic Curve: Spek 6-8 Pre-A S601 Online

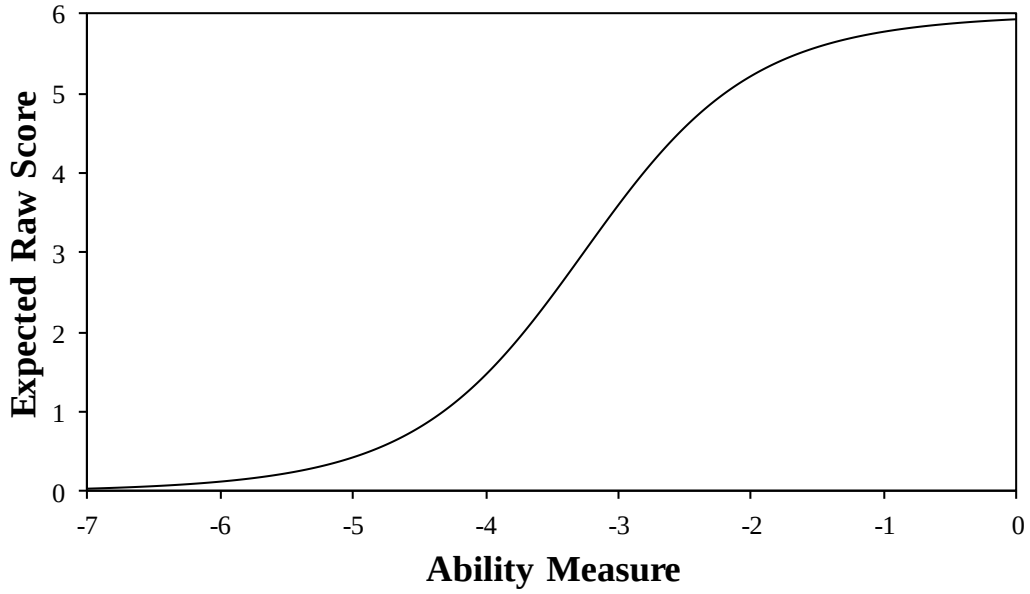


Figure 2.8.4.4.2
Test Characteristic Curve: Spek 6-8 A S601 Online

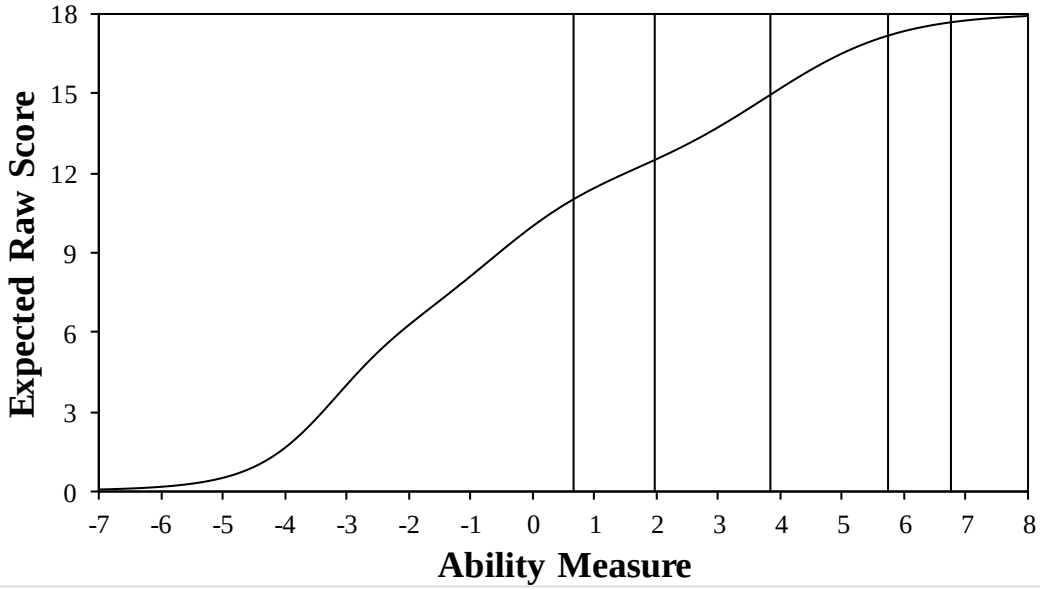
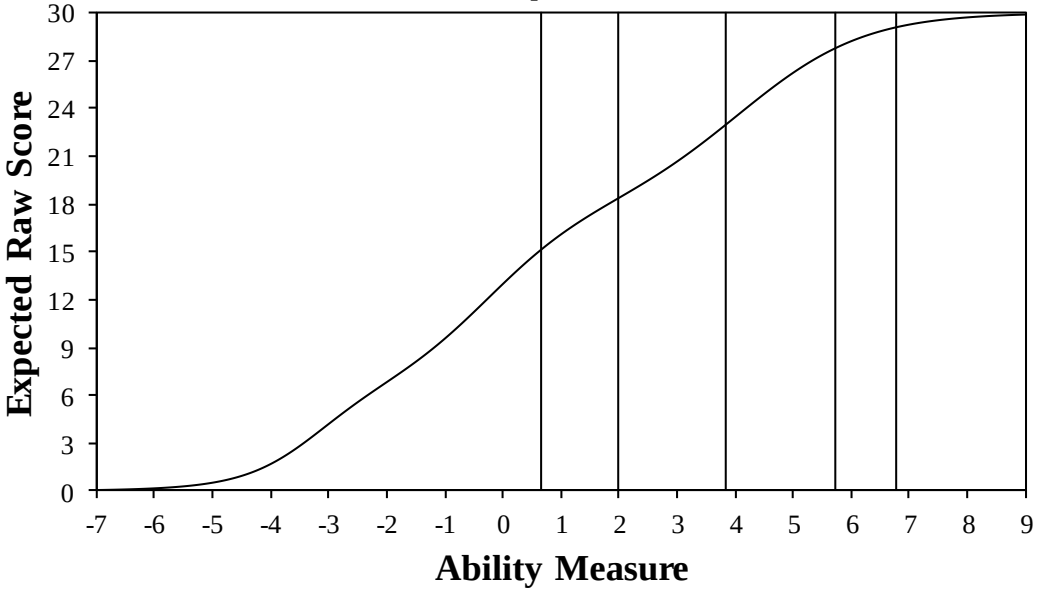


Figure 2.8.4.4.3
Test Characteristic Curve: Spek 6-8 B/C S601 Online



2.8.4.5 Grades 9-12

Figure 2.8.4.5.1
Test Characteristic Curve: Spek 9-12 Pre-A S601 Online

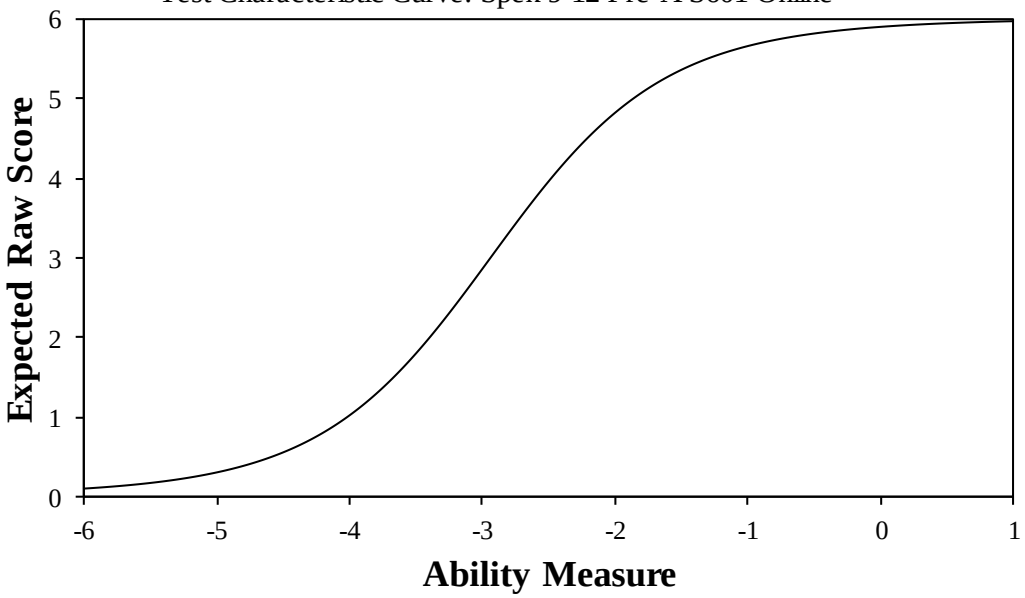


Figure 2.8.4.5.2
Test Characteristic Curve: Spek 9-12 A S601 Online

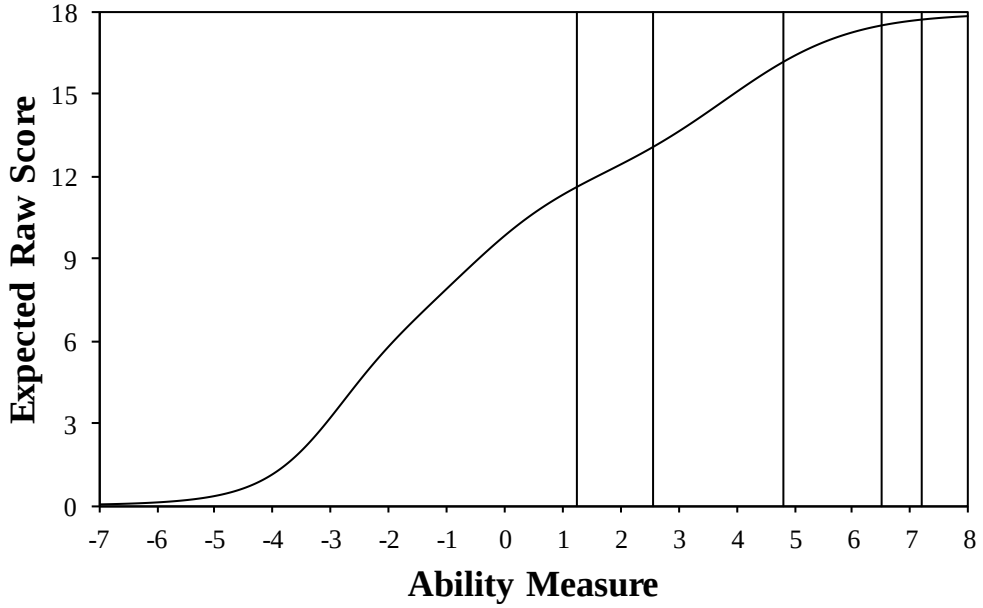
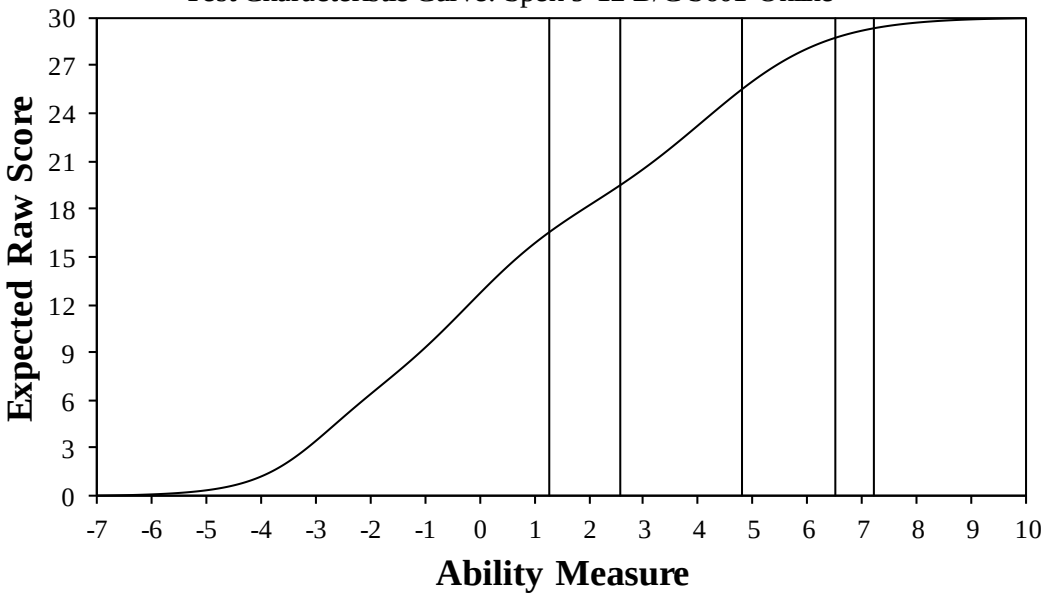


Figure 2.8.4.5.3
Test Characteristic Curve: Spek 9-12 B/C S601 Online



2.9 Test Information Function

With Rasch measurement models, as with any measurement model that is based on item response theory, one can use the item/task information function (Lord, 1980) to model the relationship between a student ability measure (in logits) and the amount of information that the students' responses to that item (or task) provides about that student's true ability. Tests perform differently for students who have differing levels of ability. Difficult items (or tasks) provide useful information for differentiating among higher-ability students but are not useful for differentiating among lower-ability students. Conversely, easy items (or tasks) provide useful information for differentiating among lower-ability students but not for differentiating among higher-ability students. Consequently, an item (or task) will provide maximum information when it is well targeted to the ability measure of the student (Reise, 1999).

The **item/task information function** indicates the amount of information that students' responses to that item (or task) provides to help reduce our uncertainty regarding a student's true ability measure. The more information we have about the ability measure, the more certain or confident we can be in that estimate of the student's ability. If the amount of information is large, that means that we have estimated with a higher degree of certainty a student whose true ability is at that level. Therefore, the ability measures for students whose scores lie within that region of the ability continuum will be reasonably close to their true values. Conversely, if the amount of information is small, that means that we have estimated with a lower degree of certainty the student whose true ability is at that level. Consequently, the ability measures for students whose scores lie within that region of the ability continuum will be further away from their true values.

Mathematically, for an item (or task), the amount of information for a given ability level is the reciprocal of the variance of the ability measure at the level. In other words, for that item (or task), the information value is the inverse squared of the standard errors of measurement for a given ability measure. Therefore, for that item (or task), the information value also provides information about the precision of the ability measure along the ability continuum.

The **test information function** (TIF) aggregates the item/task information functions across all the items (and/or tasks) on the test form or in the item pool. Since for an item (or task) the information value is the inverse squared of an ability measure's standard error of measurement,

the TIF reflects, for the whole test, the standard error of measurement for all ability measures. When the TIF is presented graphically as the test information curve, it shows how well the test is measuring across the continuum of student ability in terms of the amount of information (i.e., certainty), or the amount of measurement precision, the test provides at each ability level. The higher the curve in a particular region of the ability continuum, the more information the test provides at the ability level.

Since the TIF is the sum of all item/task information functions on the test form (Lord, 1980), the TIF depends on the information functions (Lord, 1980) of the individual items/tasks included on the test form or in the item pool. The shape of the test information curve depends on several factors, including the number and characteristics of items/tasks, the item response theory model used, and the values of the item/task parameters. With some exceptions, there is a general pattern to the shape of test information curves. Test information curves peak in the region of the student ability continuum where the test provides higher discrimination and more precise measurement as compared to other regions where the curve is less peaked, normally at the lower and upper ends of the ability continuum. When the test form consists of multiple-choice items such as in the Listening and Reading domains, the test information curve is usually unimodal.

The parameter values for the individual categories on the scoring tools that raters use to evaluate students' responses to the tasks, in addition to the factors mentioned earlier, affect the shape of the test information curves for the Writing and Speaking tests. Accordingly, some refer to these test information curves as "category information functions" (Engelhard & Wind, 2018). The scoring scales that the raters use have more score categories than the scoring schemes used for evaluating students' responses to multiple-choice items, which typically have just two categories— "right" or "wrong." Additionally, we designed the scoring scales to measure a wide range of student performance on a task. Consequently, the resulting adjacent score category boundaries may not be equidistant, and, indeed, in some cases, they may even be far apart if raters assign few scores in certain categories. In this situation, a test information curve will have one (or more) dips in the region(s) between the adjacent score category boundaries, indicating the loss of information in the corresponding ability range(s) and a decrease in the amount of information that certain score categories provide (Engelhard & Wind, 2018). Therefore, the shape of a test information curve for an ACCESS Writing or Speaking test may not be unimodal and instead may have two (or more) peaks. For example, suppose that a test information curve

reveals a dip in the region of the student writing ability continuum where raters would have assigned a score of 3. That suggests that students who received a score of 3 may have displayed potentially substantively meaningful differences in writing ability that the raters were not able to adequately distinguish when they used the 9-point Writing Scoring Scale to assign scores or, alternatively, that the score categories did not describe salient characteristics of students' writing that would make it possible for the raters to distinguish reliably among the students' responses in that region of the student ability continuum (Engelhard & Wind, 2018, pp. 316-319). The ACCESS Writing and Speaking tests are not the only assessments that have test information curves with these unusual shapes. The test information curves for other tests composed of open-ended tasks, such as the National Assessment of Educational Progress Writing assessment, also show a similar "dipping" pattern (Muraki, 1993).

The figures in this section plot the TIFs and show graphically the amount of information that the test provided across the continuum of student ability. For each test form, the five vertical lines in the figure indicate the ACCESS cut scores for the highest grade in each grade-level cluster, dividing the figure into six sections denoting the WIDA proficiency levels (PL 1– PL 6) for the domain. The test information curve and the corresponding ACCESS cut-score lines are both expressed on the ACCESS logit scale. Note that for the Speaking test, in Tier Pre-A, all scores are within the PL 1.0 range, so for some graphs there are no vertical lines showing the cut scores between proficiency levels.

Inclusion of the ACCESS cut-score lines in these figures are meant only to facilitate the visual interpretation of the test information curves relative to the ACCESS cut scores by domains. These lines provide a benchmark for WIDA and CAL assessment experts to examine the ability range for which each test seems to be more (or less) accurate in estimating student ability. Readers should note that most states that use ACCESS for ELLs do not make reclassification decisions based solely on students' domain scale scores. Rather, the majority of these states set their reclassification (or exit) criterion based on a student's Overall composite scale score, which is a weighted sum of a student's four domain scale scores. Only a few states use those four domain scale scores in addition to the student's Overall composite scale score when making a reclassification decision. Therefore, from the WIDA policy perspective, it is more important to ensure that we minimize the measurement error near the cut score that most states use to set their

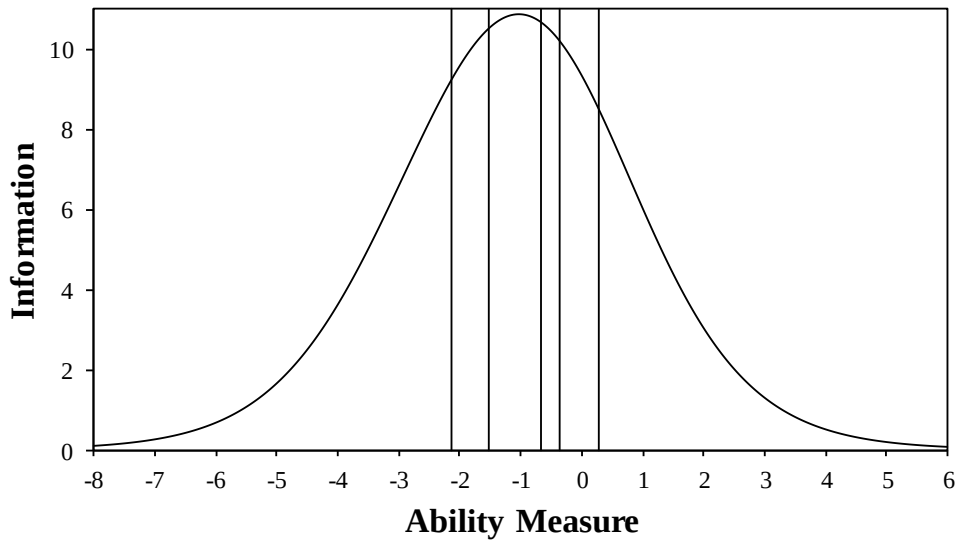
reclassification criterion on the Overall composite scale score. We report the conditional standard errors of measurement (CSEMs) for the Overall composite scale scores in Section 5.6.

In addition to the TIF graphs by tier, for the Writing and Speaking tests, in the same graph we provide plots of the TIFs across tiers, by grade-level cluster. Test users may find it useful to compare the ability ranges across tiers, noting for each tier where the curve displays its highest peaks (i.e., where the most measurement information is provided). For example, as shown in Figure 2.9.3.1.3, the test information curve across tiers for Writing Grade 1 reveals that the Writing Grade 1 Tier A form provided more information about student ability measures that were either just below the PL 2 cut score or just below the PL 5 cut score. By contrast, the Writing Grade 1 Tier B/C form provided more information about the student ability measures that were either just above the PL 2 cut score or just above the PL 5 cut score. The plot also shows that the Writing Grade 1 Tier A form provided more information for those student ability measures in the lowest range (i.e., ability measures of -0.5 logits or lower), while the Writing Grade 1 Tier B/C form provided more information than the Grade 1 Tier A form for the rest of the student ability measures, especially those in the higher ability range. Lastly, consistent with the purposes of the test design, there is also considerable overlap between the ranges of writing ability that the two forms cover.

2.9.1 Listening

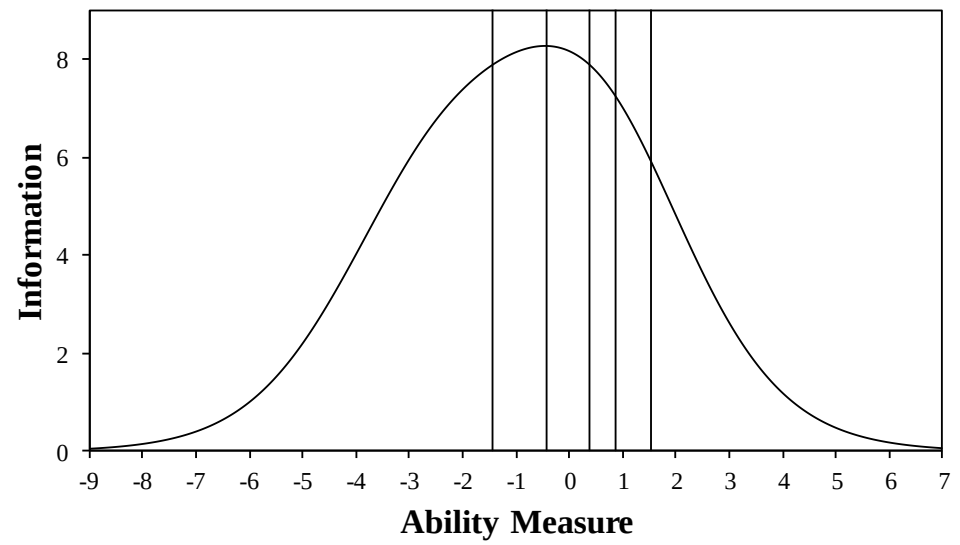
2.9.1.1 Grade 1

Figure 2.9.1.1
Test Information Curve: List 1 S601 Online



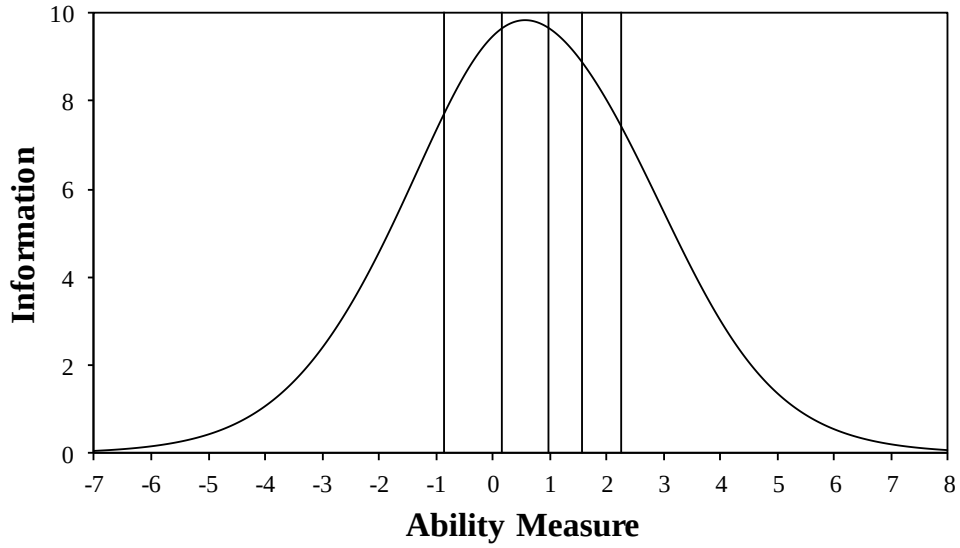
2.9.1.2 Grade 2-3

Figure 2.9.1.2
Test Information Curve: List 2-3 S601 Online



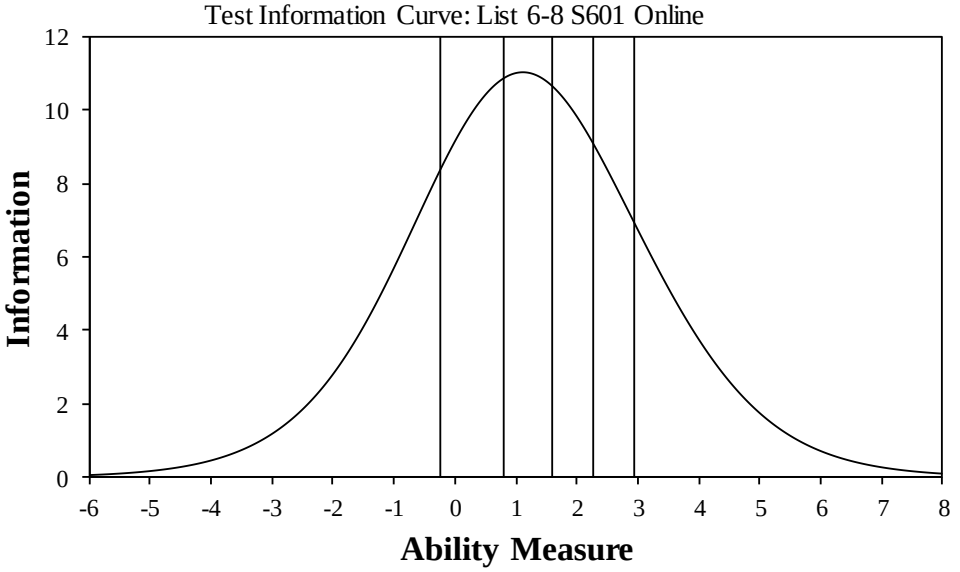
2.9.1.3 Grades 4-5

Figure 2.9.1.3
Test Information Curve: List 4-5 S601 Online



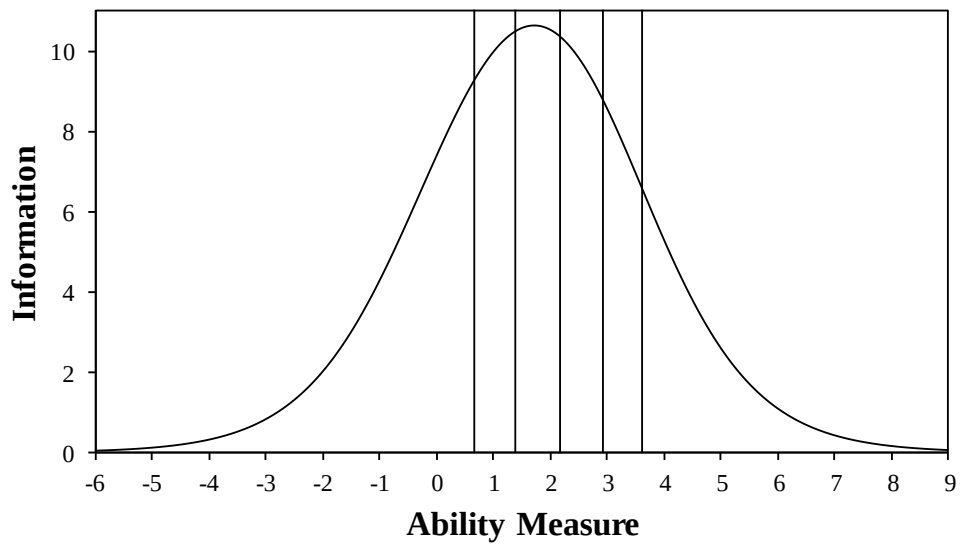
2.9.1.4 Grades 6-8

Figure 2.9.1.4
Test Information Curve: List 6-8 S601 Online



2.9.1.5 Grades 9-12

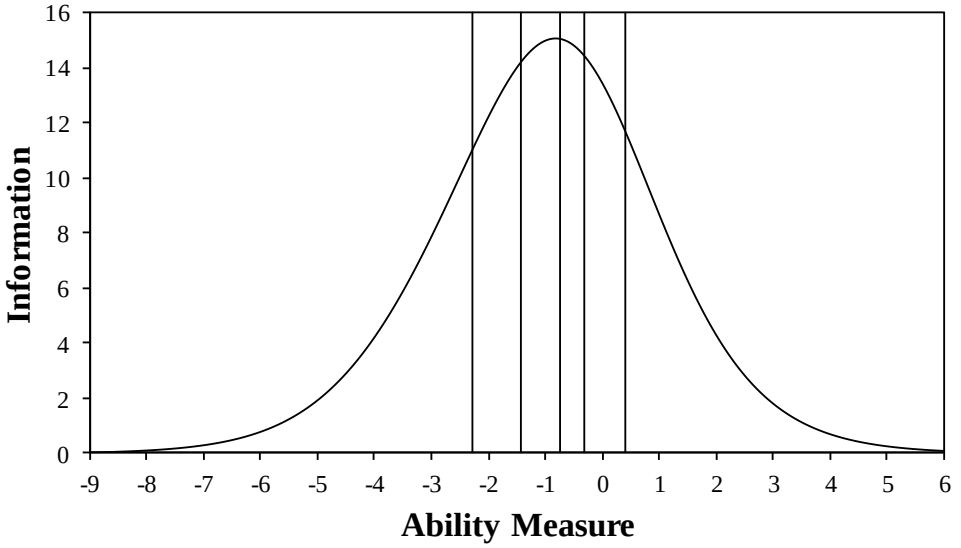
Figure 2.9.1.5
Test Information Curve: List 9-12 S601 Online



2.9.2 Reading

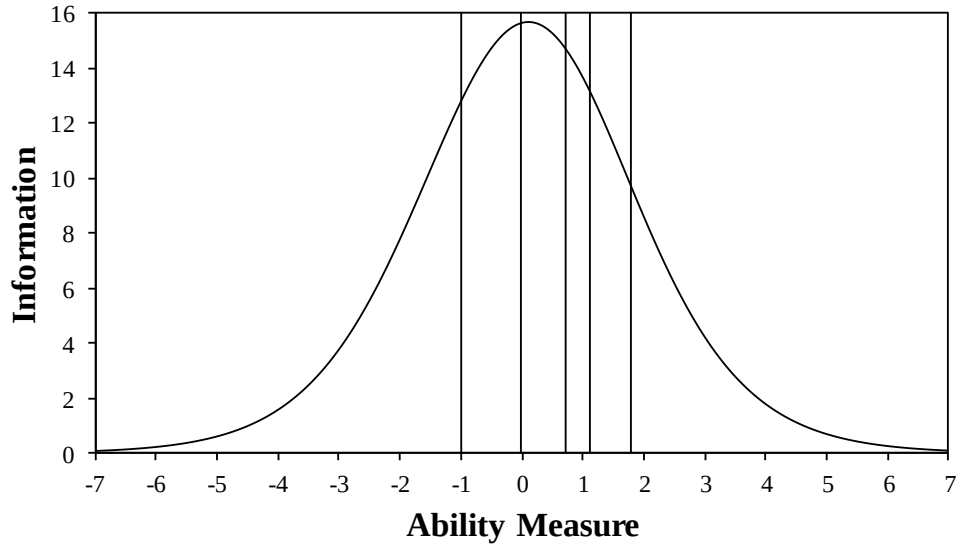
2.9.2.1 Grade 1

Figure 2.9.2.1
Test Information Curve: Read 1 S601 Online



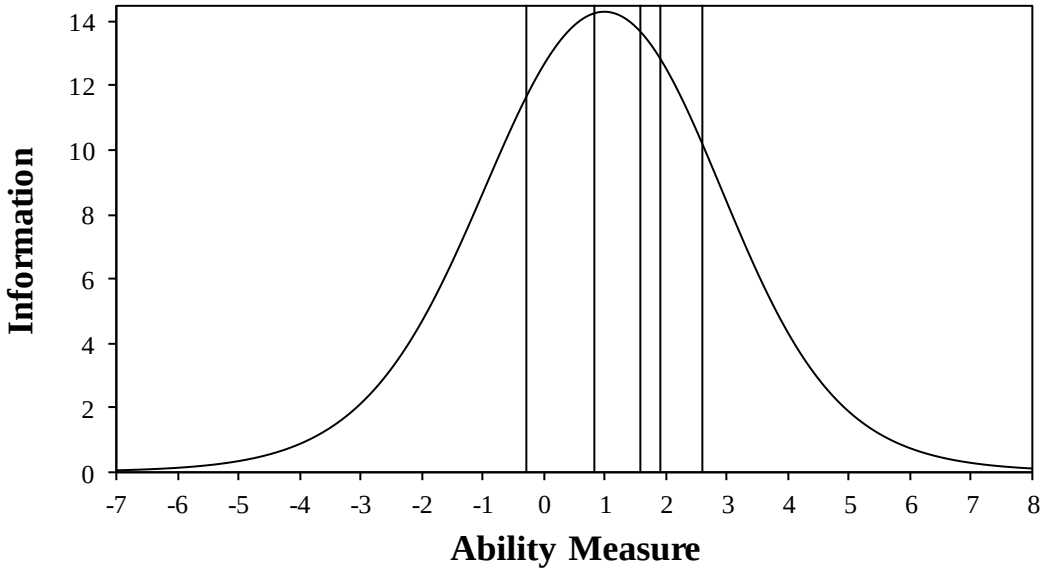
2.9.2.2 Grade 2-3

Figure 2.9.2.2
Test Information Curve: Read 2-3 S601 Online



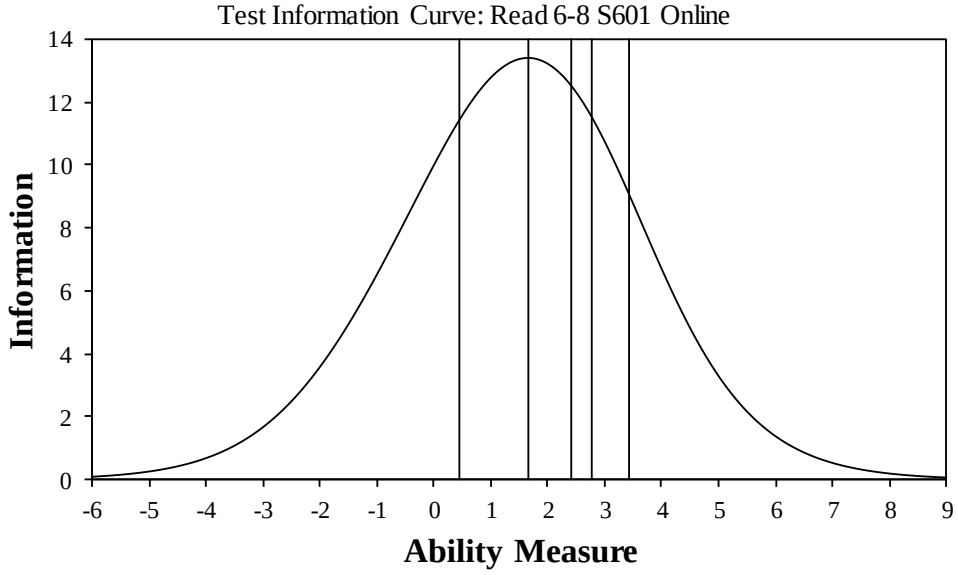
2.9.2.3 Grades 4-5

Figure 2.9.2.3
Test Information Curve: Read 4-5 S601 Online



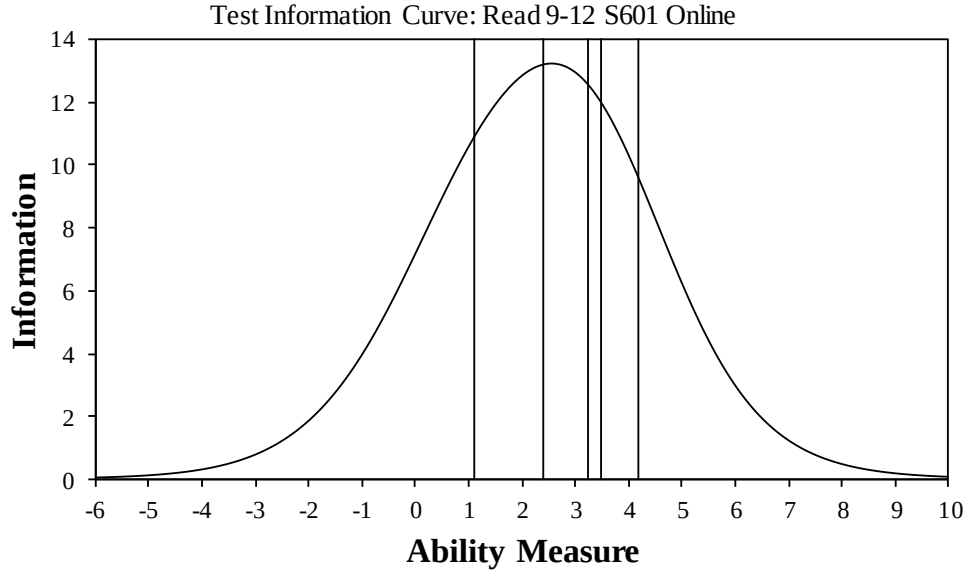
2.9.2.4 Grades 6-8

Figure 2.9.2.4
Test Information Curve: Read 6-8 S601 Online



2.9.2.5 Grades 9-12

Figure 2.9.2.5
Test Information Curve: Read 9-12 S601 Online



2.9.3 Writing

2.9.3.1 Grade 1

Figure 2.9.3.1.1 Test Information Curve: Writ 1 A S601 Online

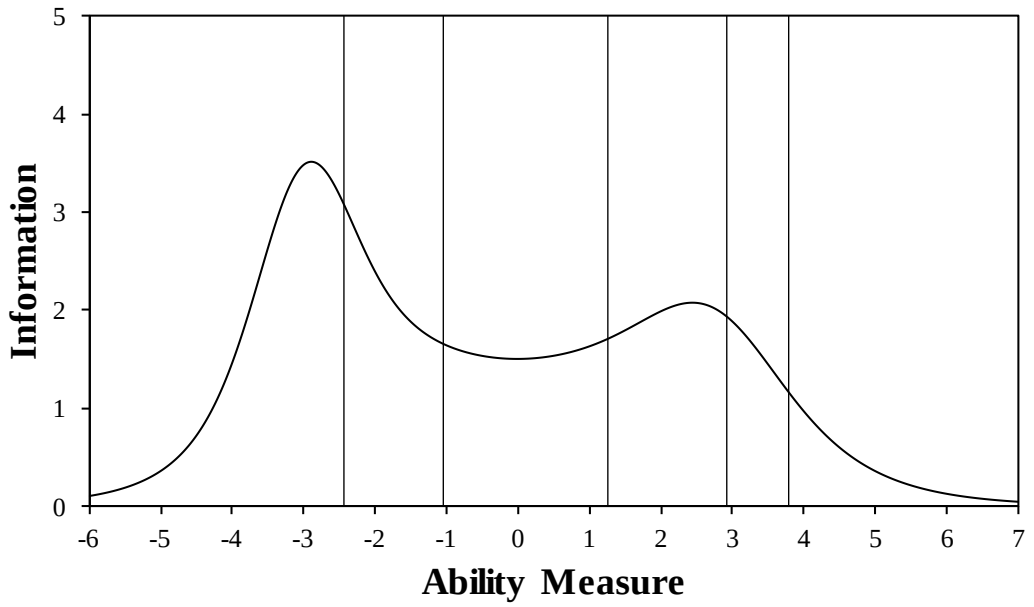


Figure 2.9.3.1.2 Test Information Curve: Writ 1 B/C S601 Online

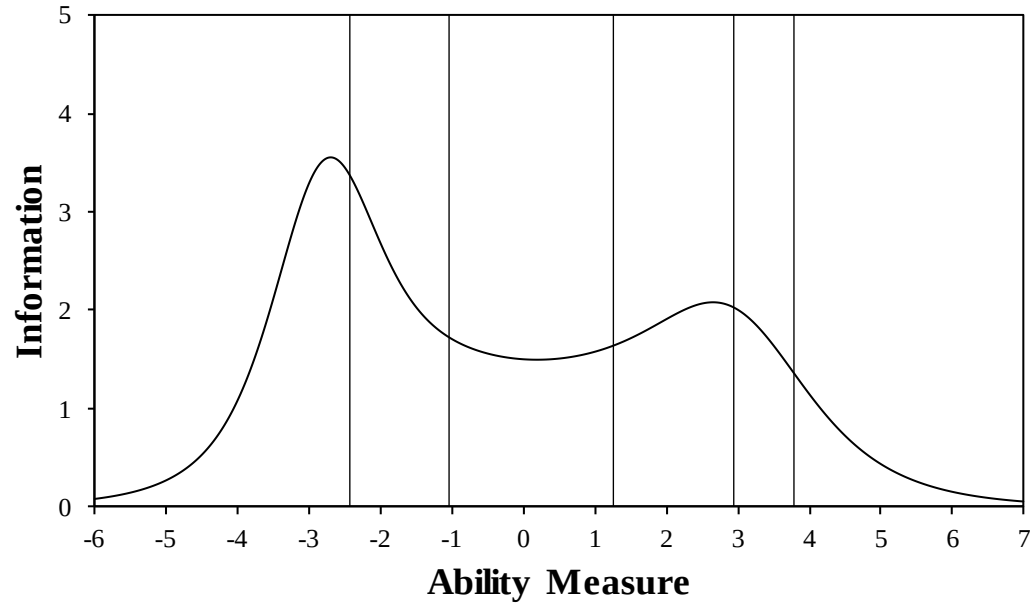
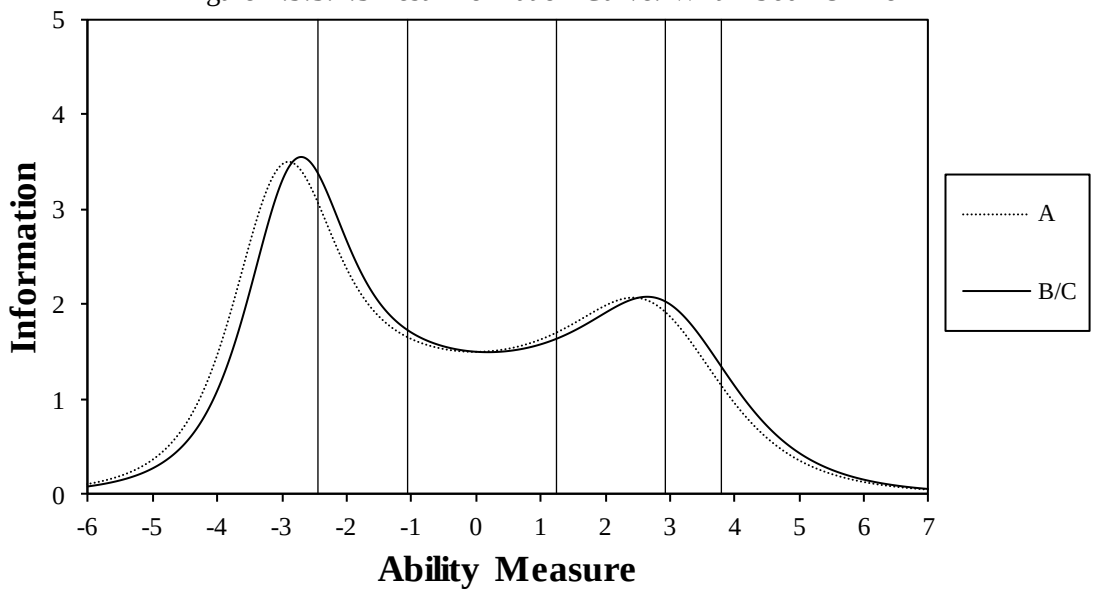


Figure 2.9.3.1.3 Test Information Curve: Writ 1 S601 Online



2.9.3.2 Grade 2-3

Figure 2.9.3.2.1 Test Information Curve: Writ 2-3 A S601 Online

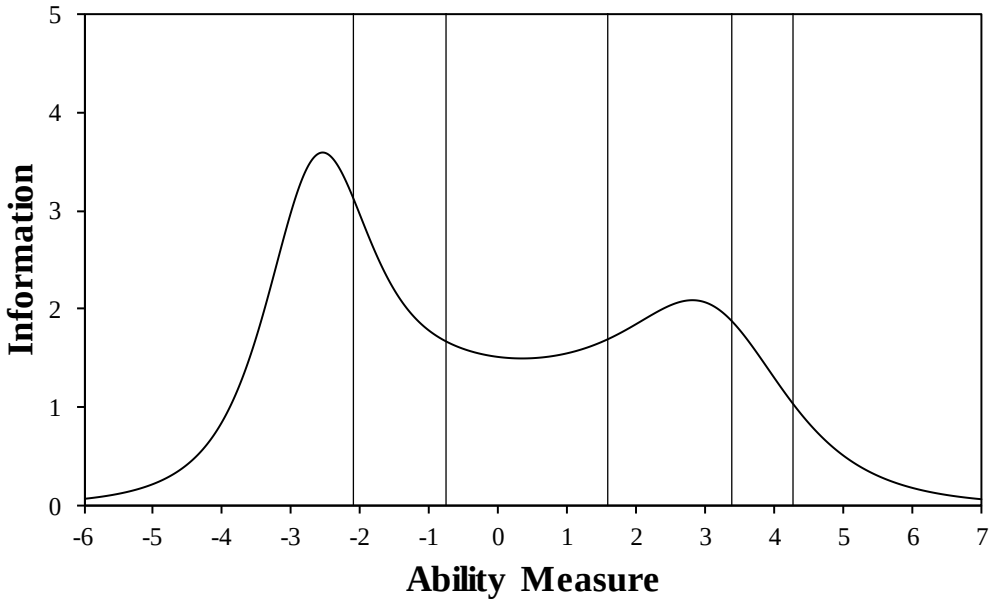


Figure 2.9.3.2.2 Test Information Curve: Writ 2-3 B/C S601 Online

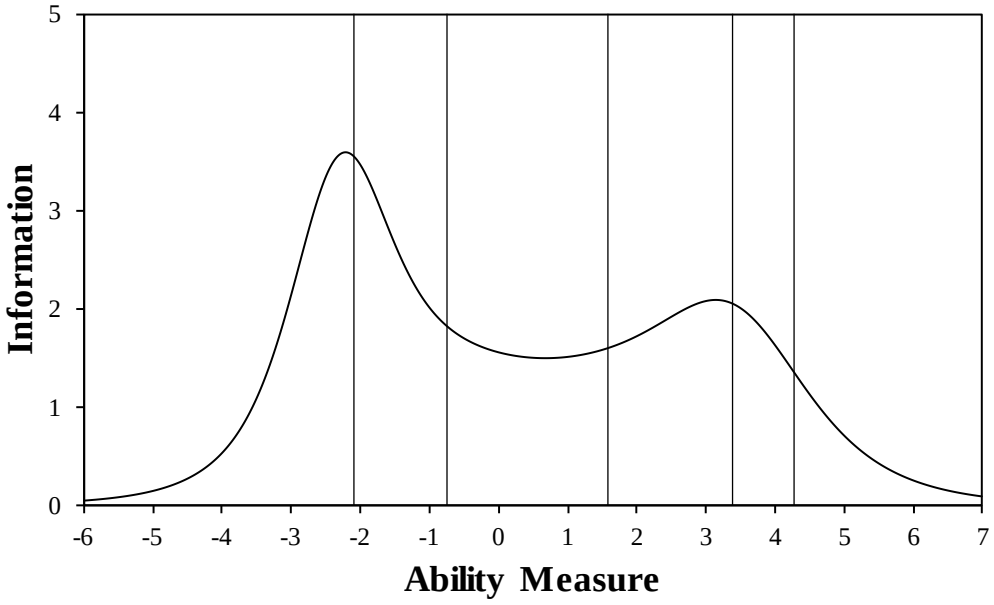
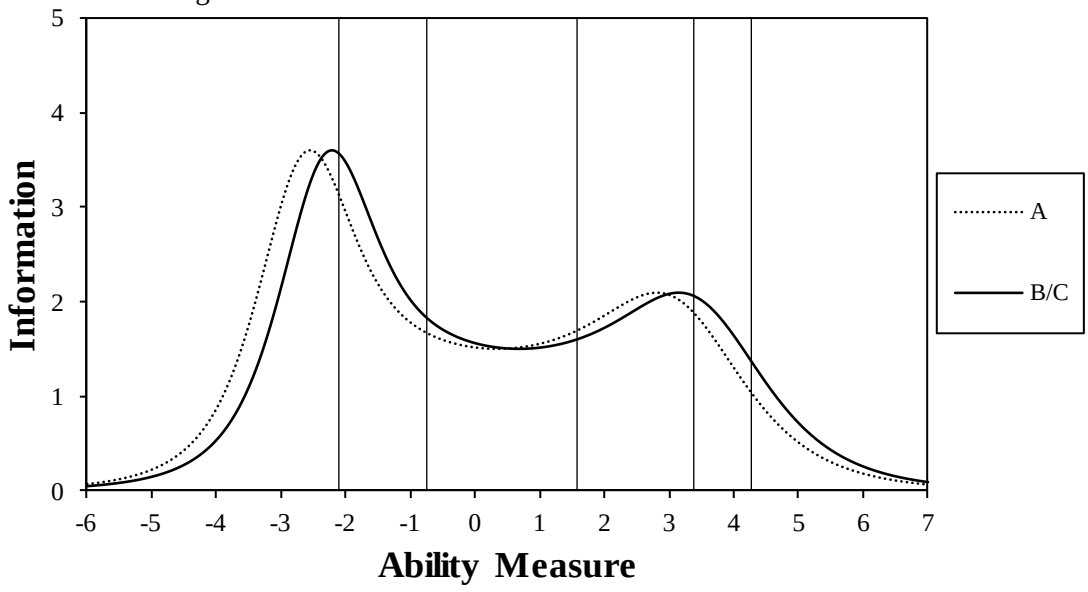


Figure 2.9.3.2.3 Test Information Curve: Writ 2-3 S601 Online



2.9.3.3 Grades 4-5

Figure 2.9.3.3.1 Test Information Curve: Writ 4-5 A S601 Online

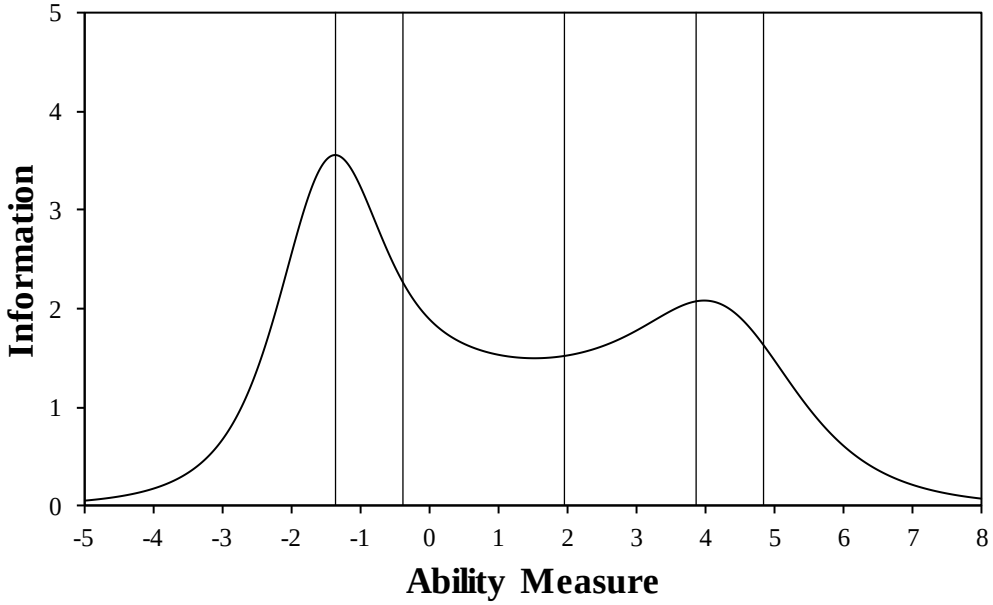


Figure 2.9.3.3.2 Test Information Curve: Writ 4-5 B/C S601 Online

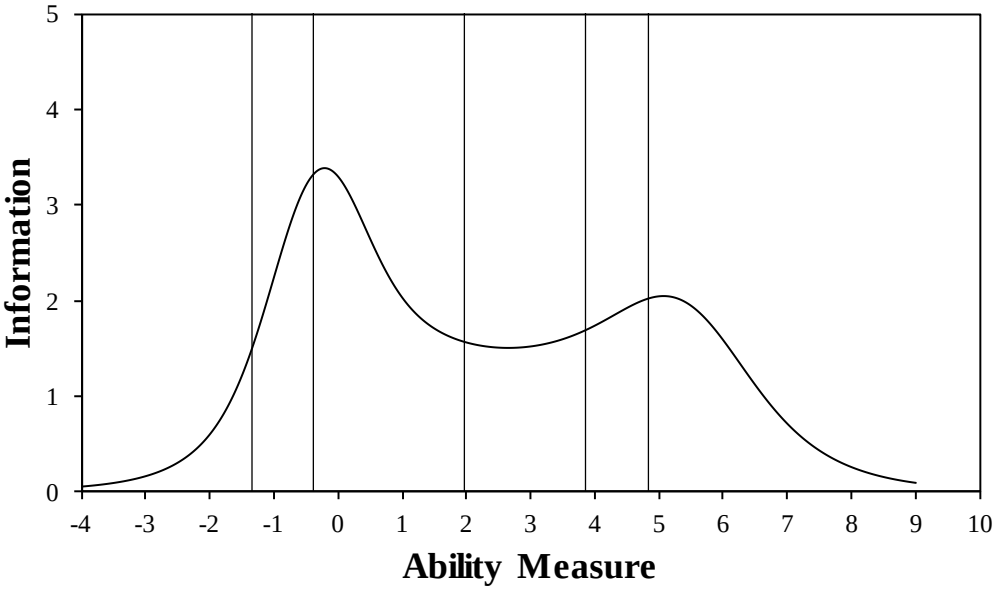
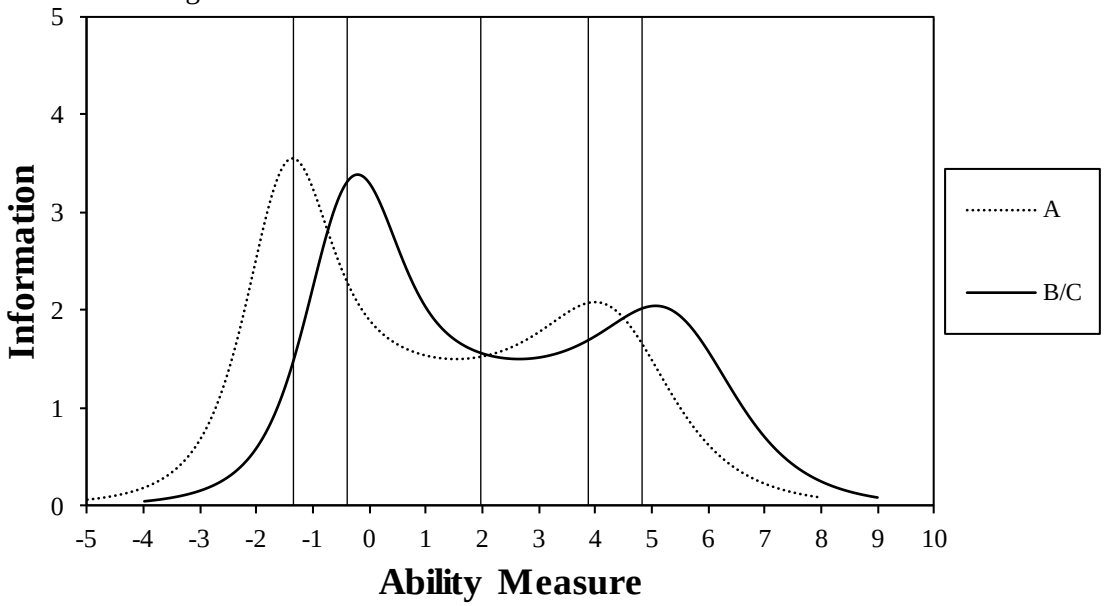


Figure 2.9.3.3.3 Test Information Curve: Writ 4-5 S601 Online



2.9.3.4 Grades 6-8

Figure 2.9.3.4.1 Test Information Curve: Writ 6-8 A S601 Online

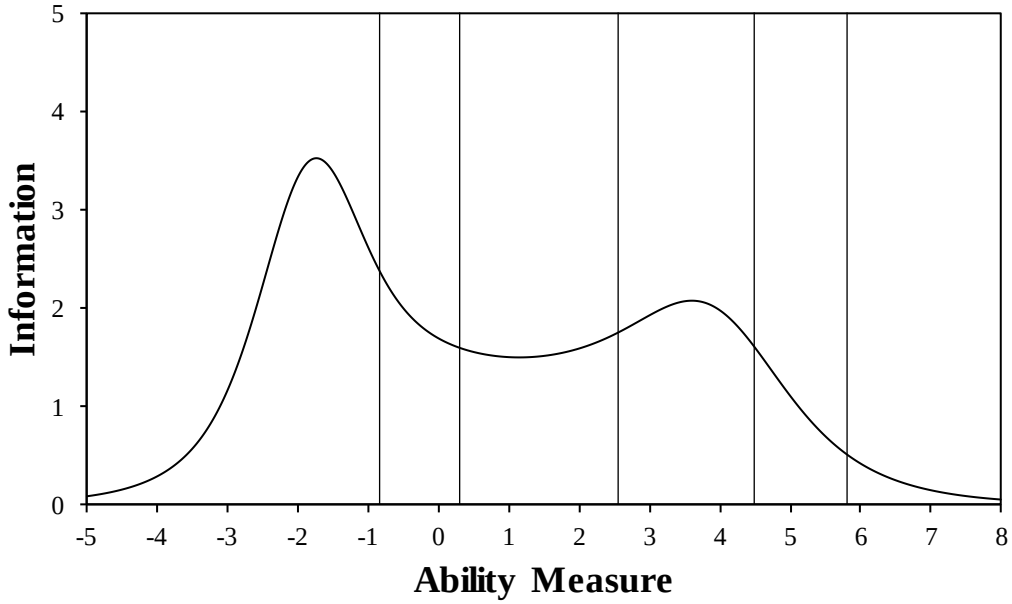


Figure 2.9.3.4.2 Test Information Curve: Writ 6-8 B/C S601 Online

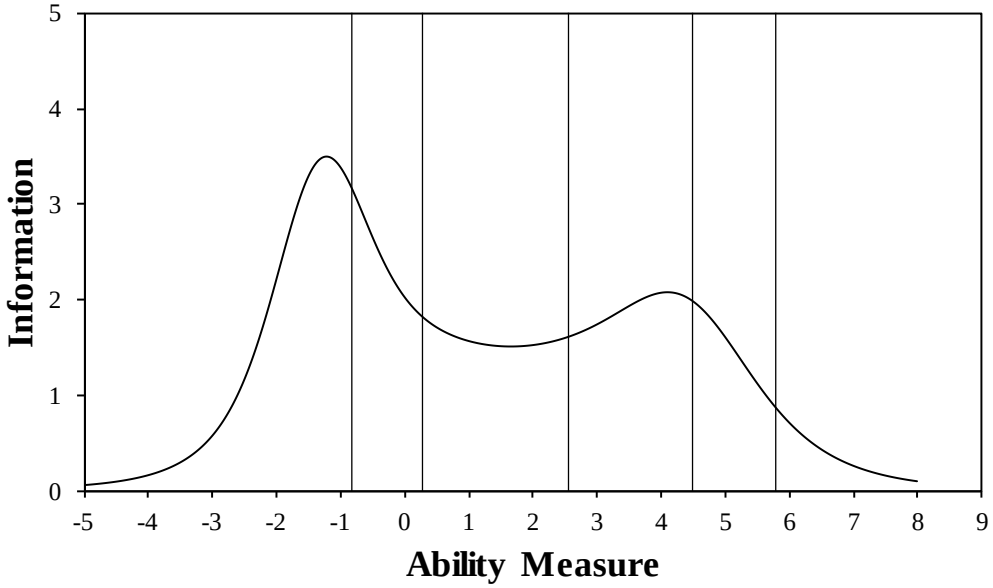
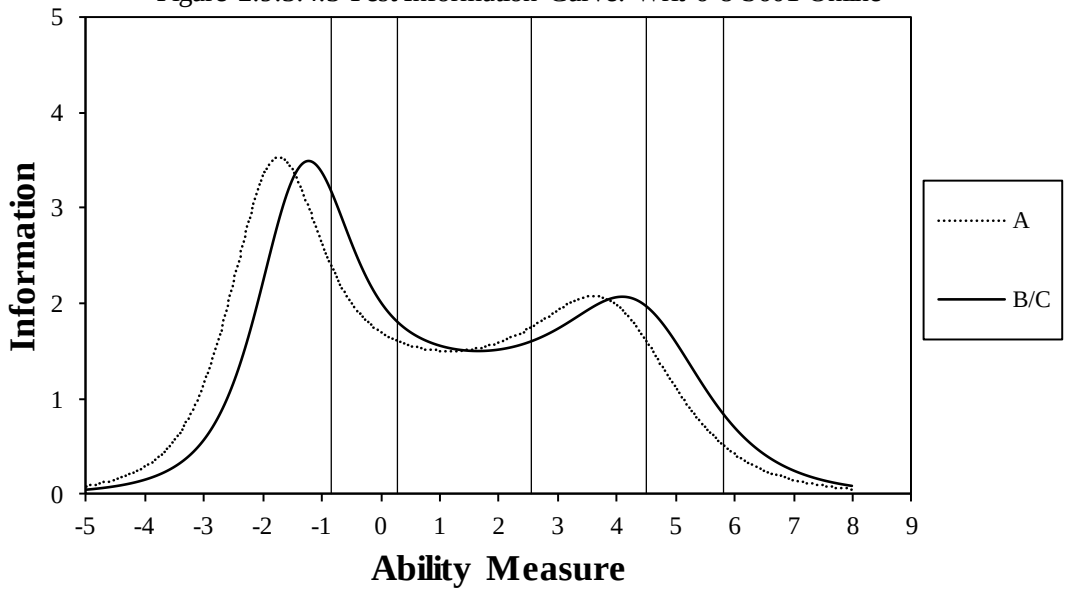


Figure 2.9.3.4.3 Test Information Curve: Writ 6-8 S601 Online



2.9.3.5 Grades 9-12

Figure 2.9.3.5.1 Test Information Curve: Writ 9-12 A S601 Online

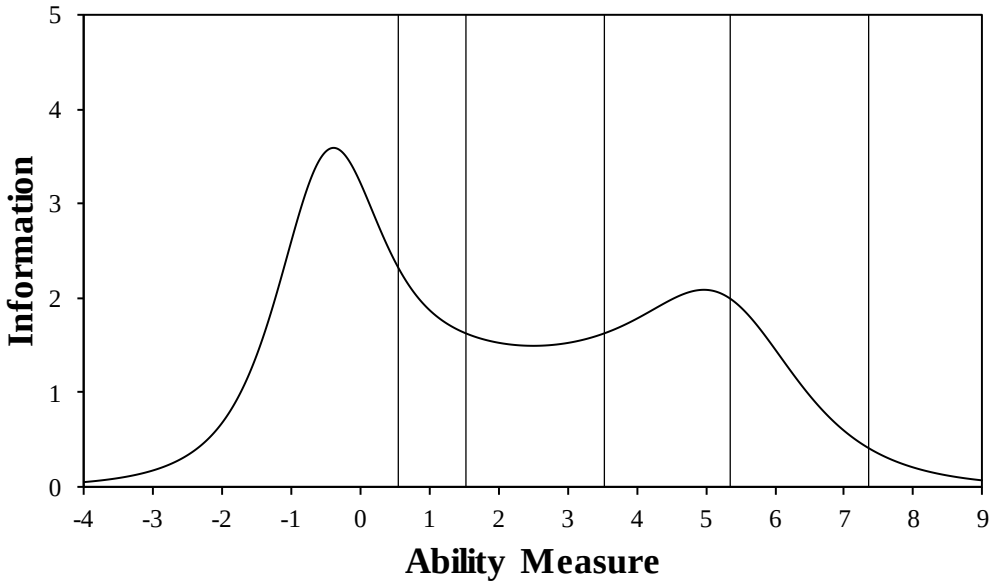
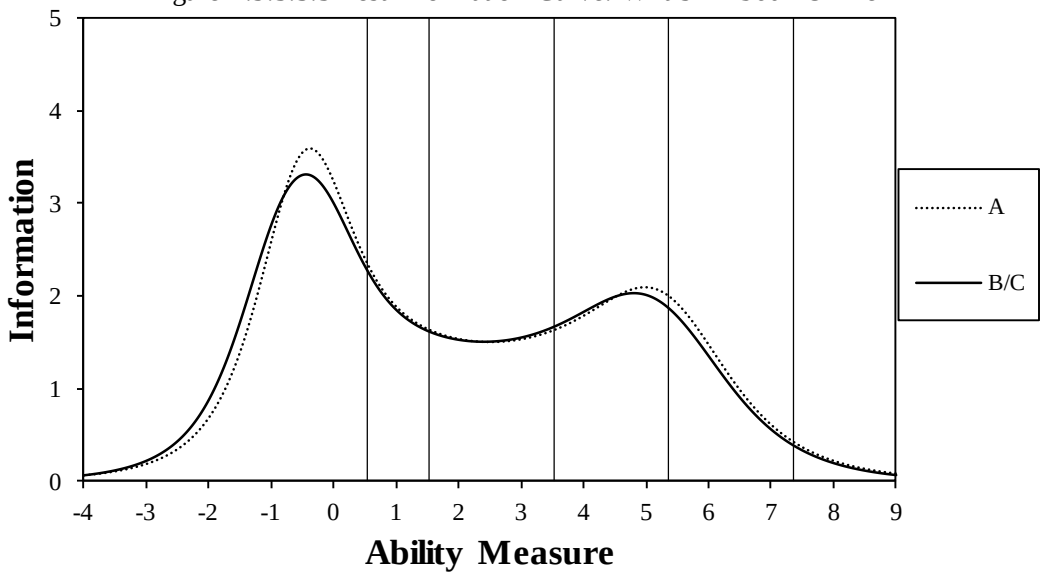


Figure 2.9.3.5.2 Test Information Curve: Writ 9-12 B/C S601 Online

Figure 2.9.3.5.3 Test Information Curve: Writ 9-12 S601 Online



2.9.4 Speaking

2.9.4.1 Grade 1

Figure 2.9.4.1.1
Test Information Curve: Spek 1 Pre-A S601 Online

Figure 2.9.4.1.2
Test Information Curve: Spek 1 A S601 Online

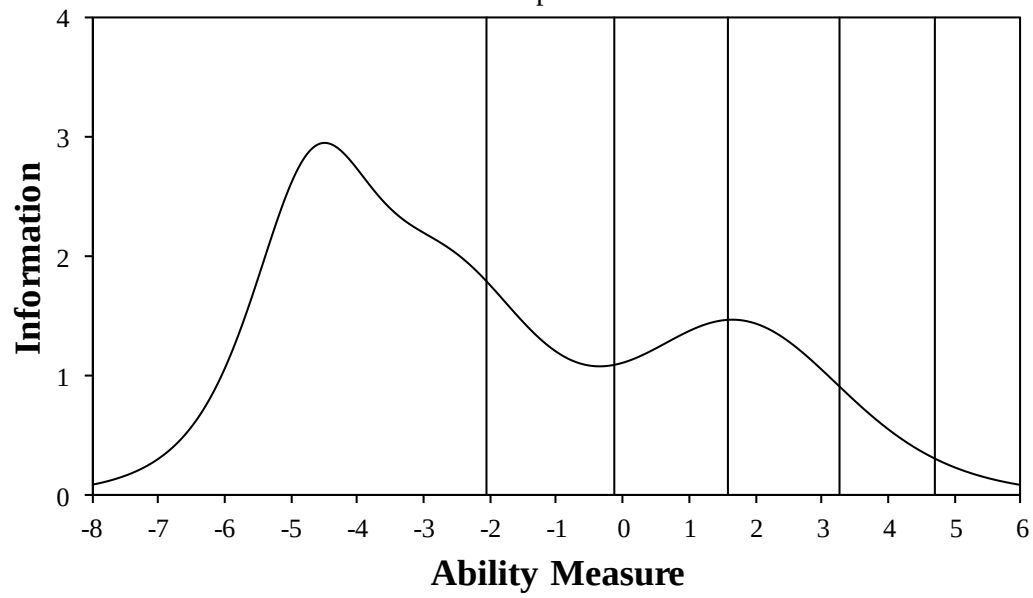


Figure 2.9.4.1.3
Test Information Curve: Spek 1 B/C S601 Online

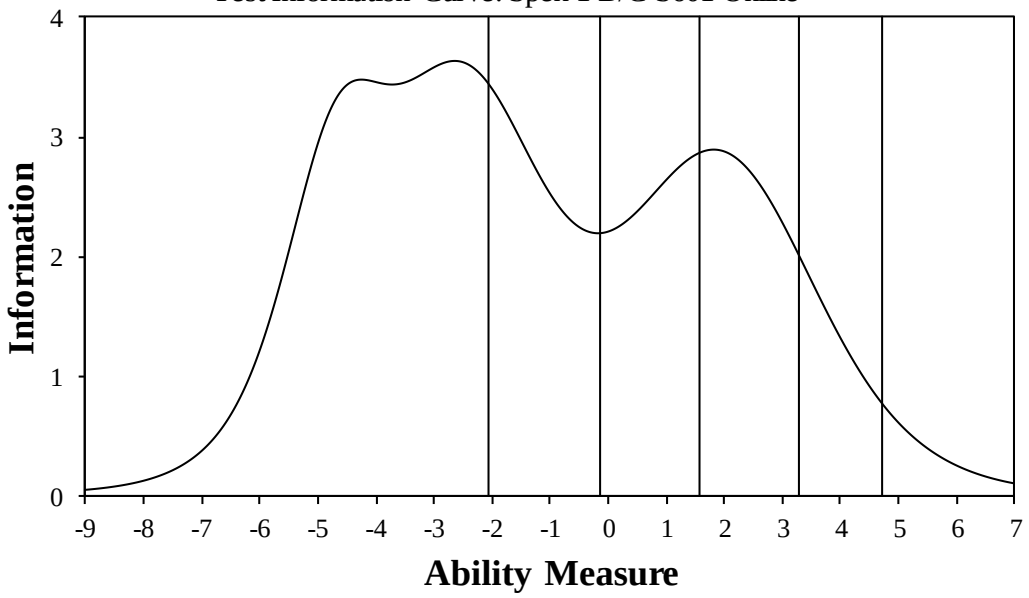
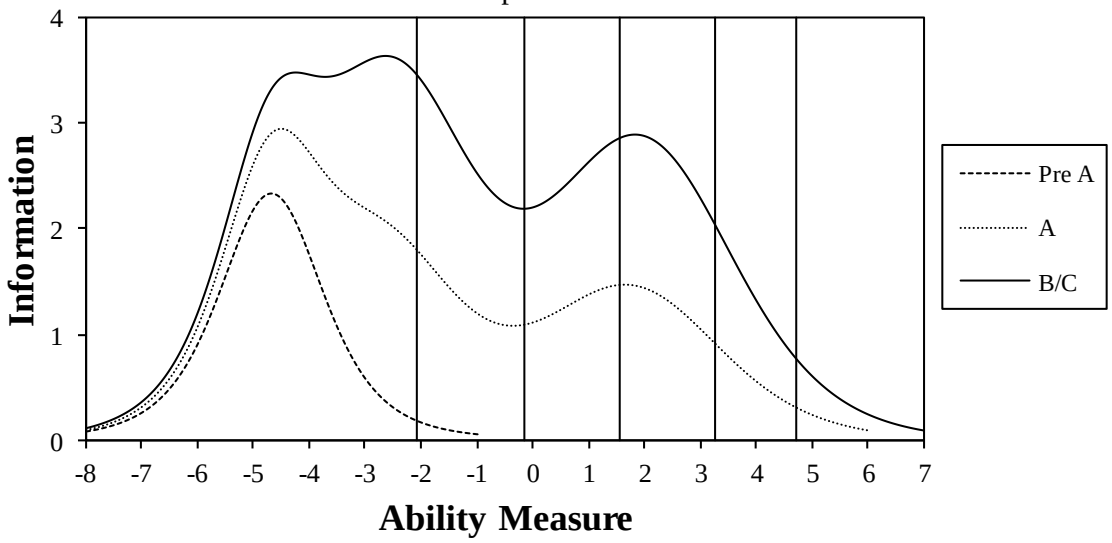


Figure 2.9.4.1.4
Test Information Curve: Spek 1 S601 Online



2.9.4.2 Grade 2-3

Figure 2.9.4.2.1
Test Information Curve: Spek 2-3 Pre-A S601 Online

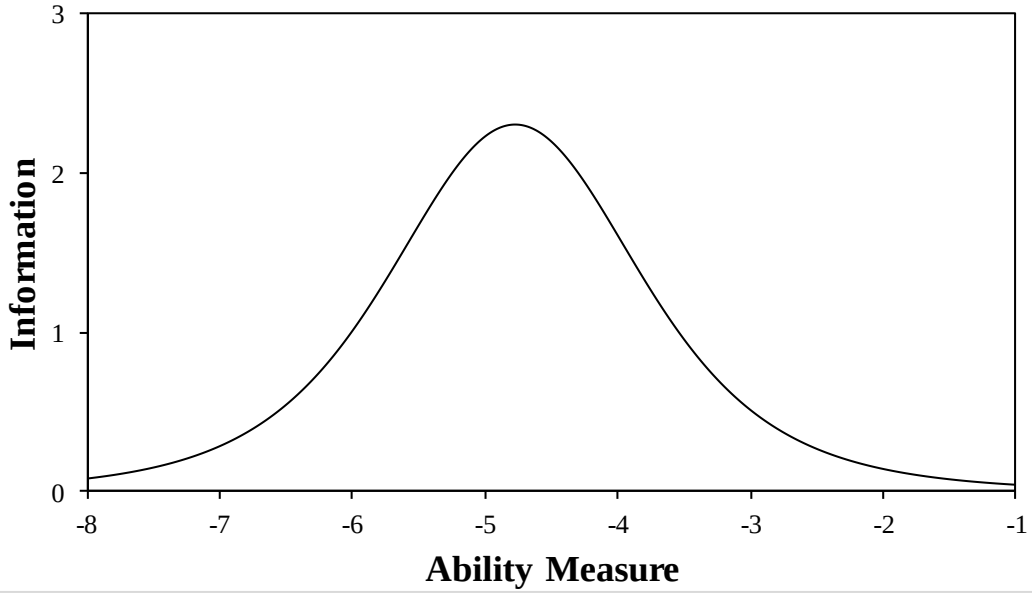


Figure 2.9.4.2.2
Test Information Curve: Spek 2-3 A S601 Online

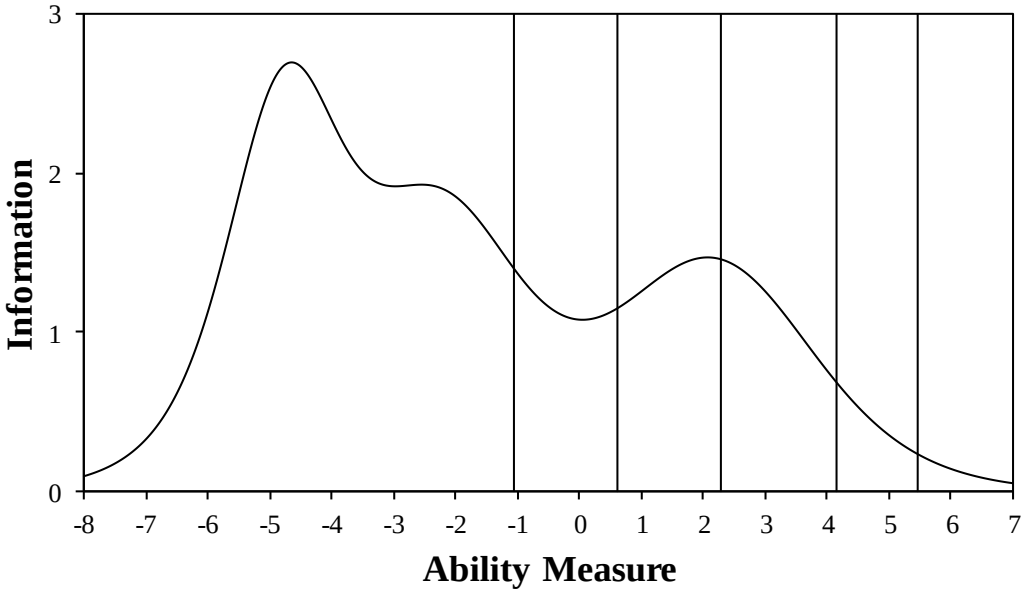


Figure 2.9.4.2.3
Test Information Curve: Spek 2-3 B/C S601 Online

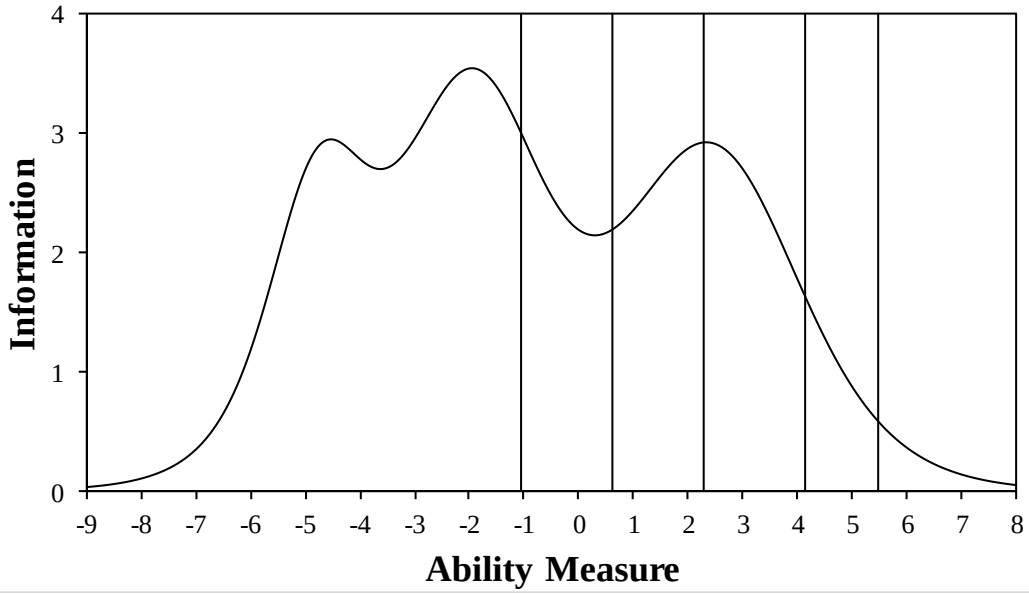
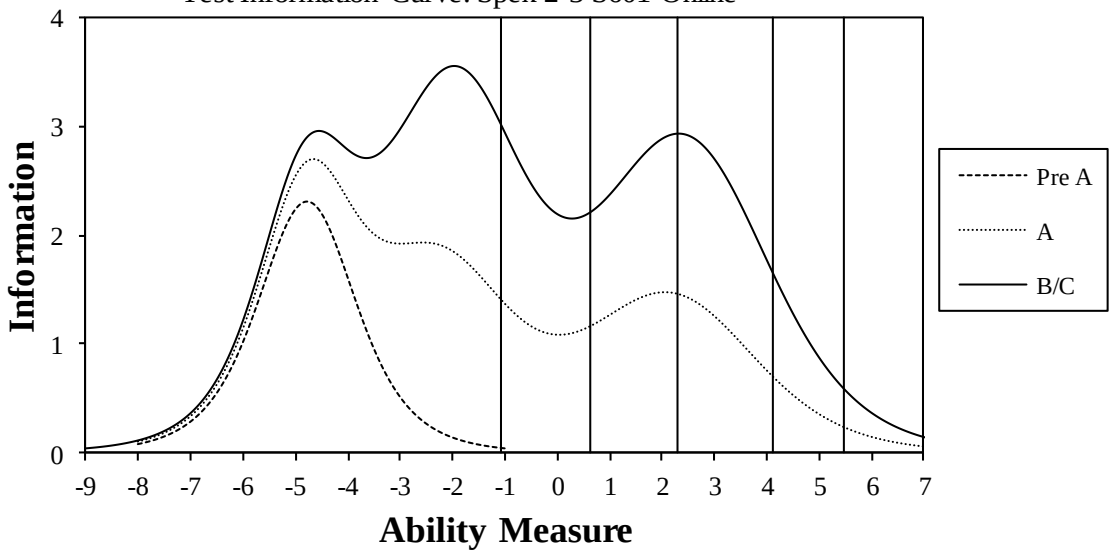


Figure 2.9.4.2.4
Test Information Curve: Spek 2-3 S601 Online



2.9.4.3 Grades 4-5

Figure 2.9.4.3.1
Test Information Curve: Spek 4-5 Pre-A S601 Online

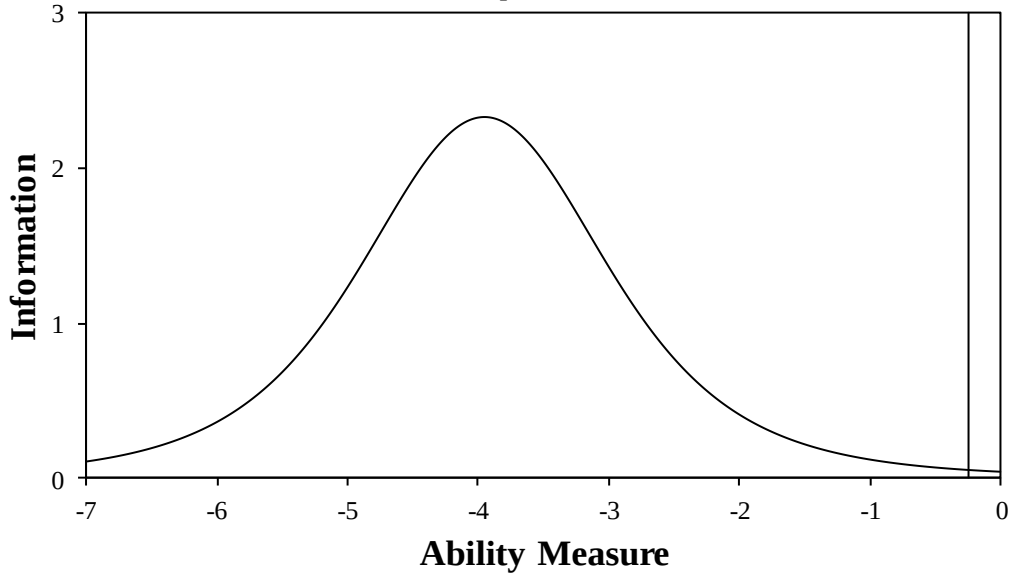


Figure 2.9.4.3.2
Test Information Curve: Spek 4-5 A S601 Online

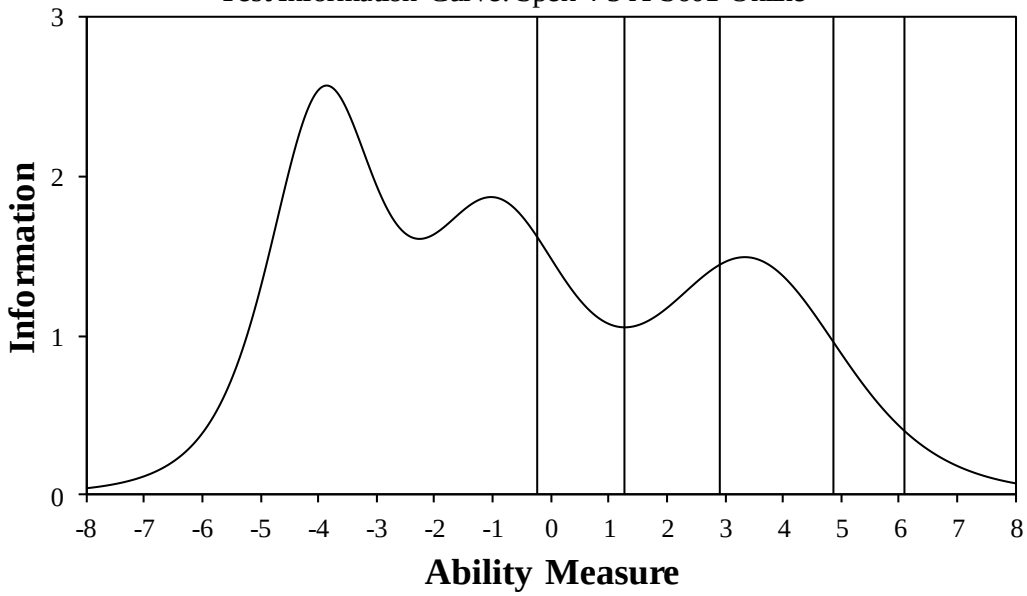


Figure 2.9.4.3.3
Test Information Curve: Spek 4-5 B/C S601 Online

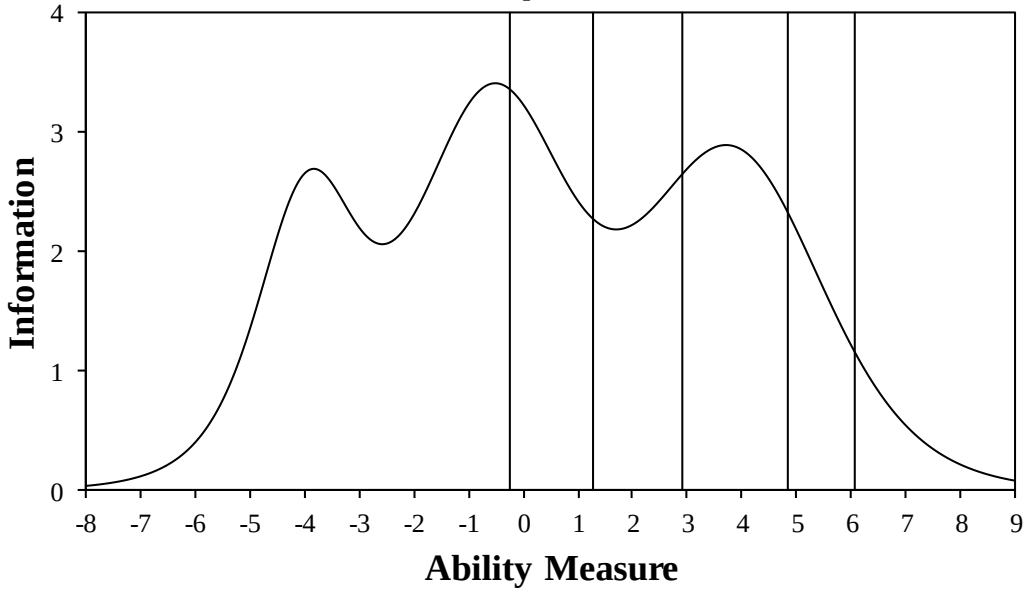
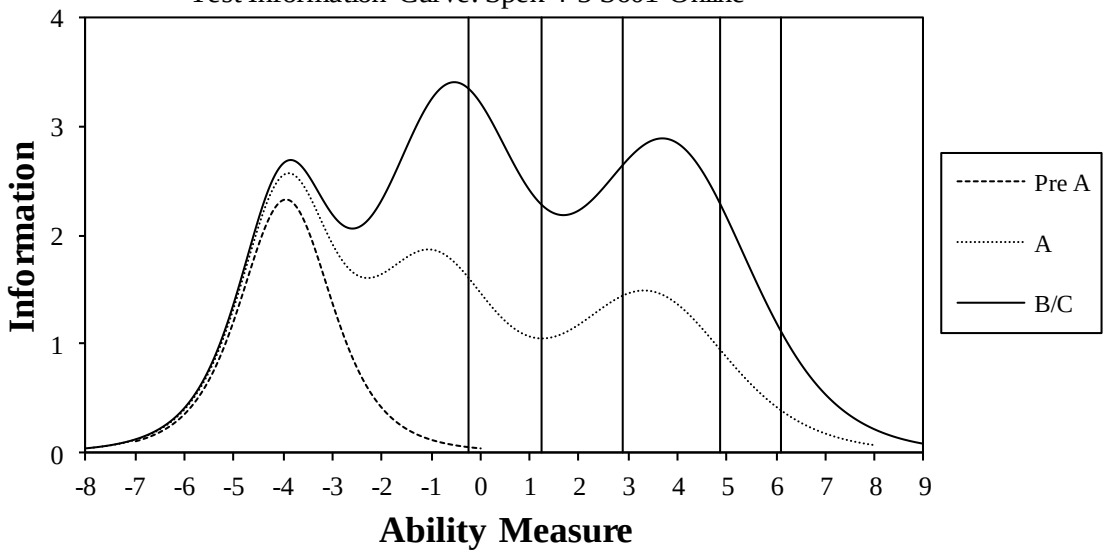


Figure 2.9.4.3.4
Test Information Curve: Spek 4-5 S601 Online



2.9.4.4 Grades 6-8

Figure 2.9.4.4.1
Test Information Curve: Spek 6-8 Pre-A S601 Online

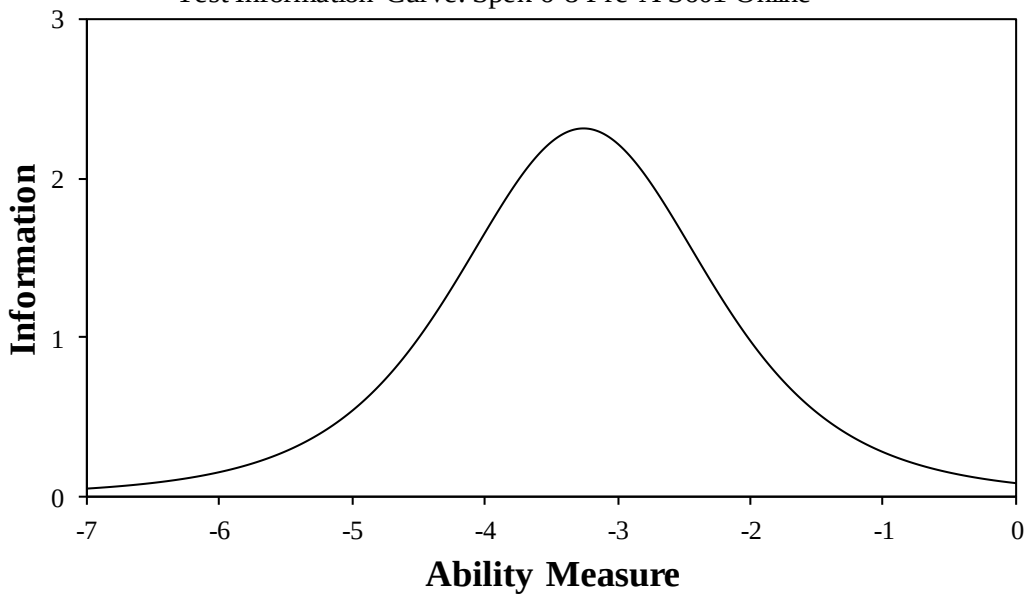


Figure 2.9.4.4.2
Test Information Curve: Spek 6-8 A S601 Online

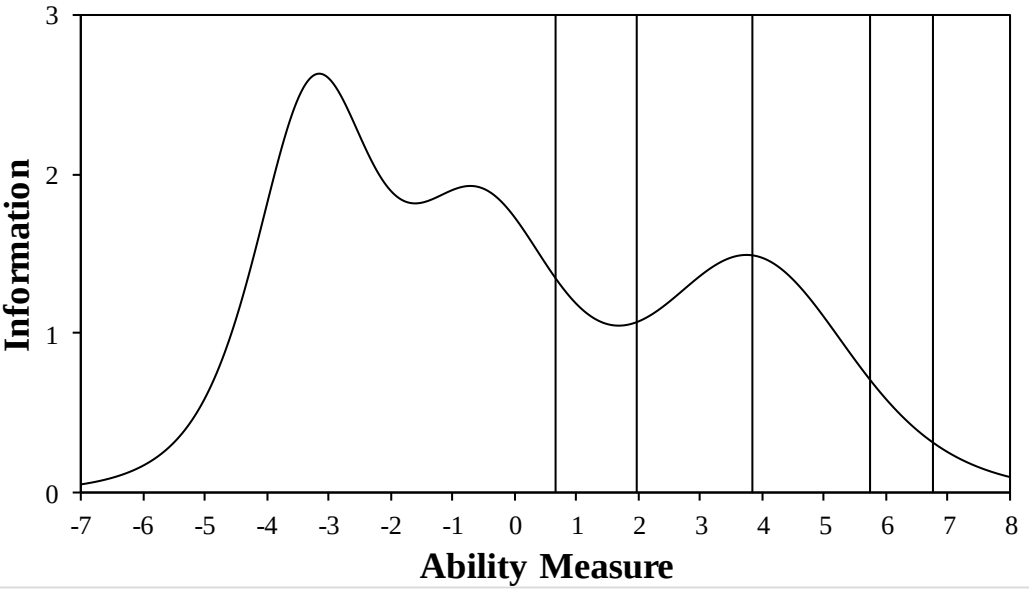


Figure 2.9.4.4.3
Test Information Curve: Spek 6-8 B/C S601 Online

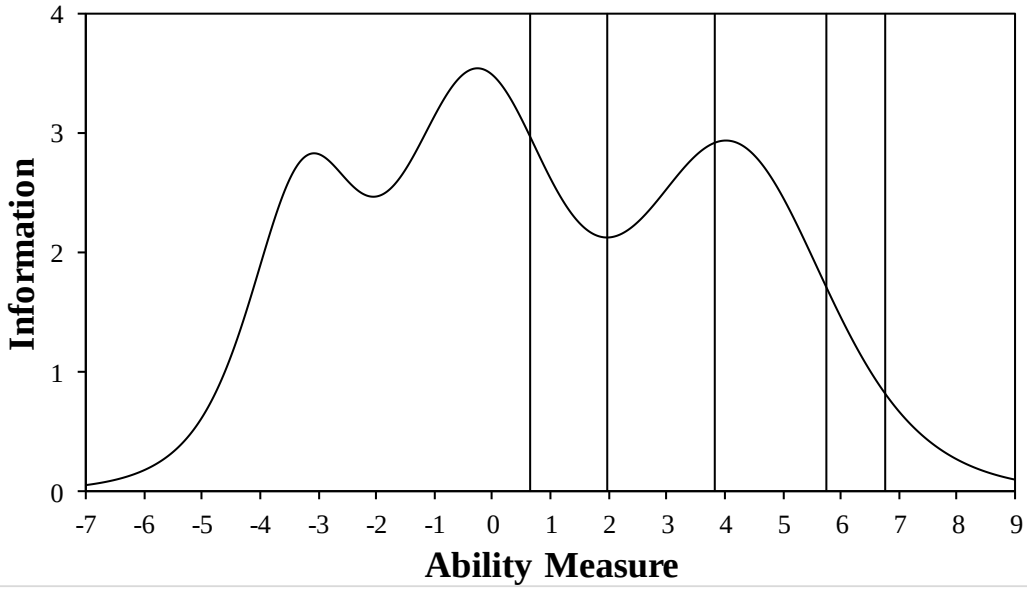
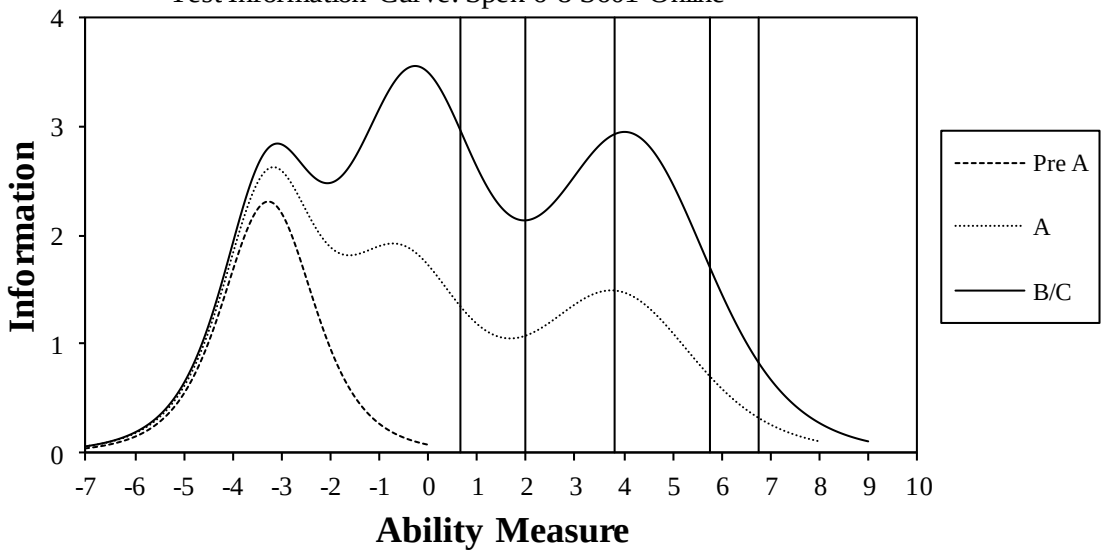


Figure 2.9.4.4.4
Test Information Curve: Spek 6-8 S601 Online



2.9.4.5 Grades 9-12

Figure 2.9.4.5.1
Test Information Curve: Spek 9-12 Pre-A S601 Online

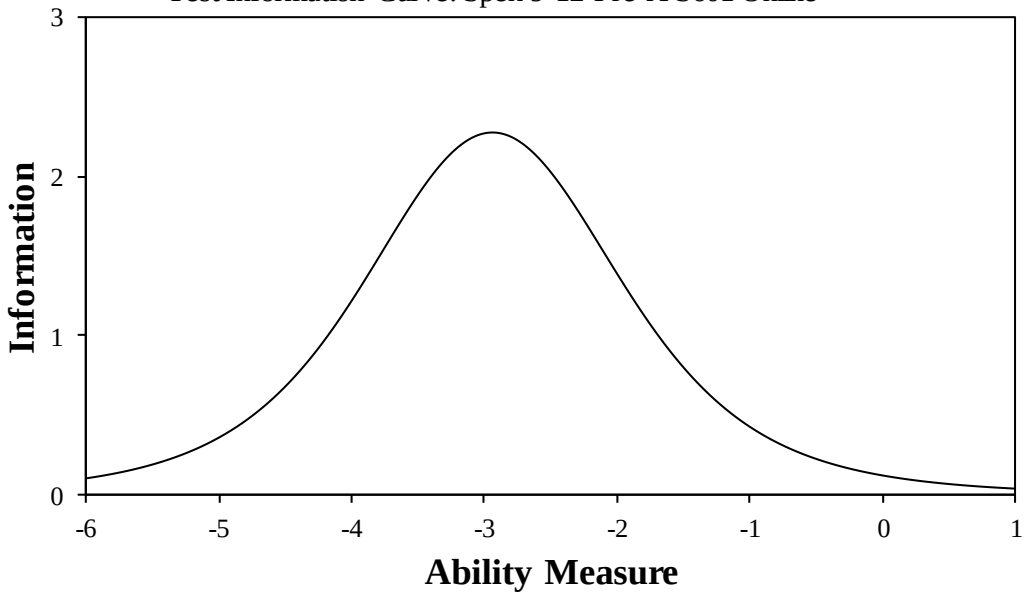


Figure 2.9.4.5.2
Test Information Curve: Spek 9-12 A S601 Online

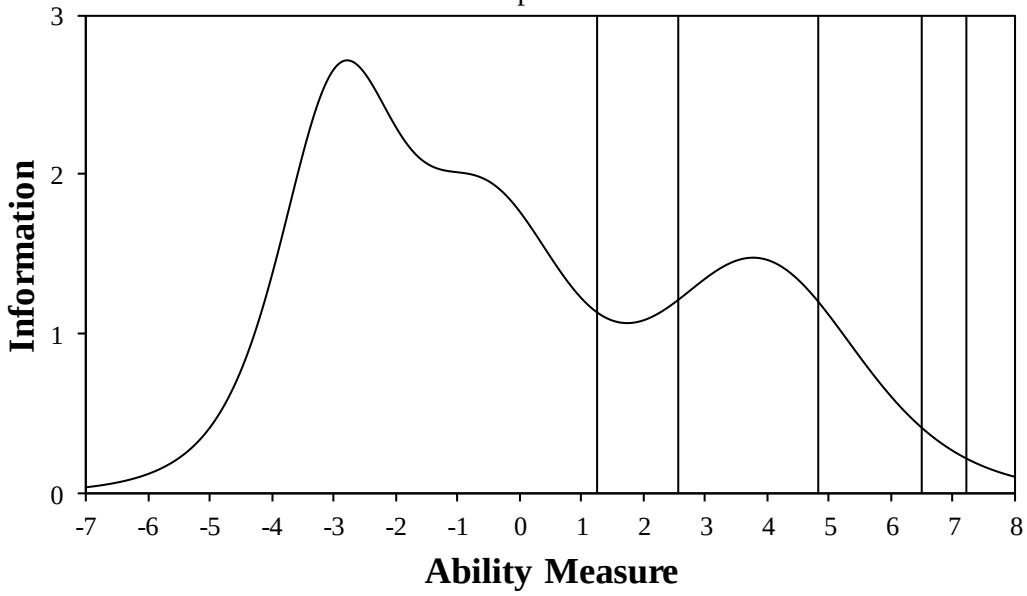


Figure 2.9.4.5.3
Test Information Curve: Spek 9-12 B/C S601 Online

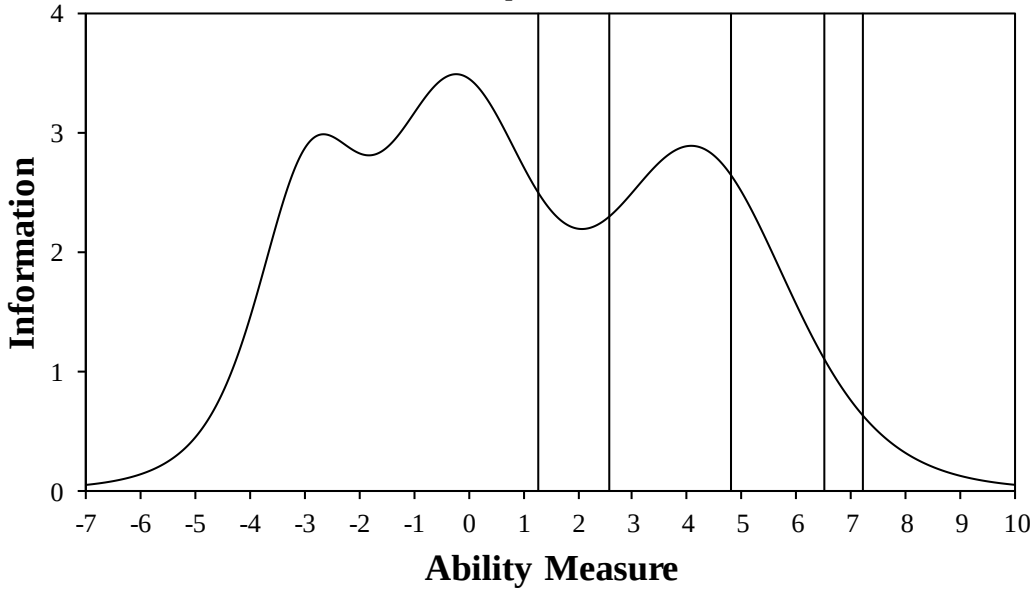
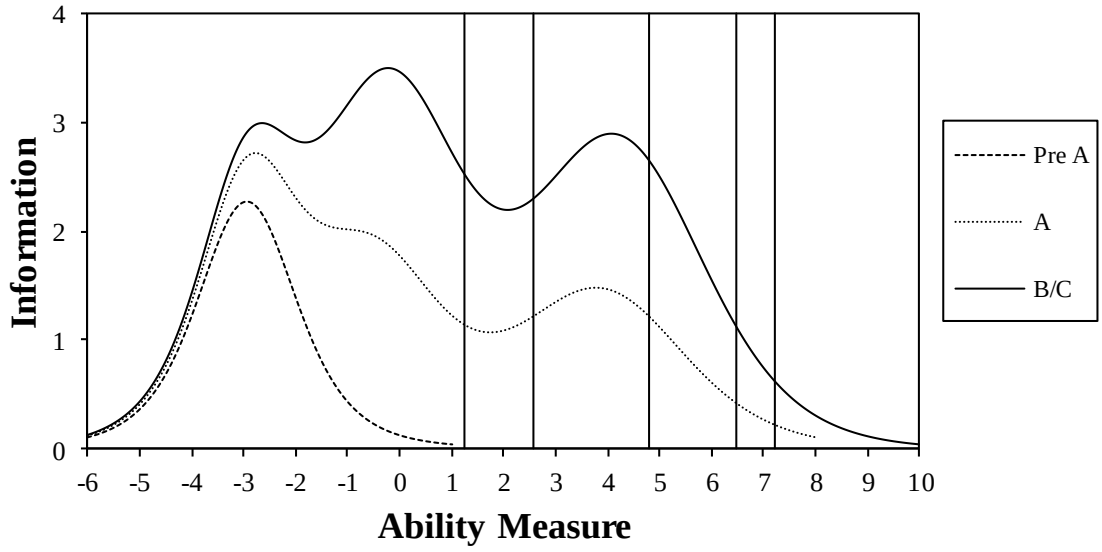


Figure 2.9.4.5.4
Test Information Curve: Spek 9-12 S601 Online



3. Analyses of Composite Scores

We calculate four composite scores for ACCESS Online: Oral Language, Literacy, Comprehension, and Overall. We calculate these composite scores as weighted averages of domain scale scores, as follows:

- Oral Language: 50% Listening + 50% Speaking
- Literacy: 50% Reading + 50% Writing
- Comprehension: 30% Listening + 70% Reading
- Overall Composite: 15% Listening + 15% Speaking + 35% Reading + 35% Writing

A policy decision by the WIDA Board, made before the first operational administration of ACCESS, resulted in the weighting, and is based on the view that literacy skills are paramount in developing academic language proficiency.

3.1 Scale Score Distribution for Composites

Figures and tables in this section provide scale score distributions for each of the composites, for each grade-level cluster.

For each cluster, the figure shows the distribution of the scale scores for the composite. We plotted the scale scores, grouped into units of five scale score points (e.g., 100–104, 105–109, 110–114, etc.), on the horizontal axis and the number of students with scale scores falling into each range on the vertical axis.

Each table shows, by grade and by total for the grade-level cluster:

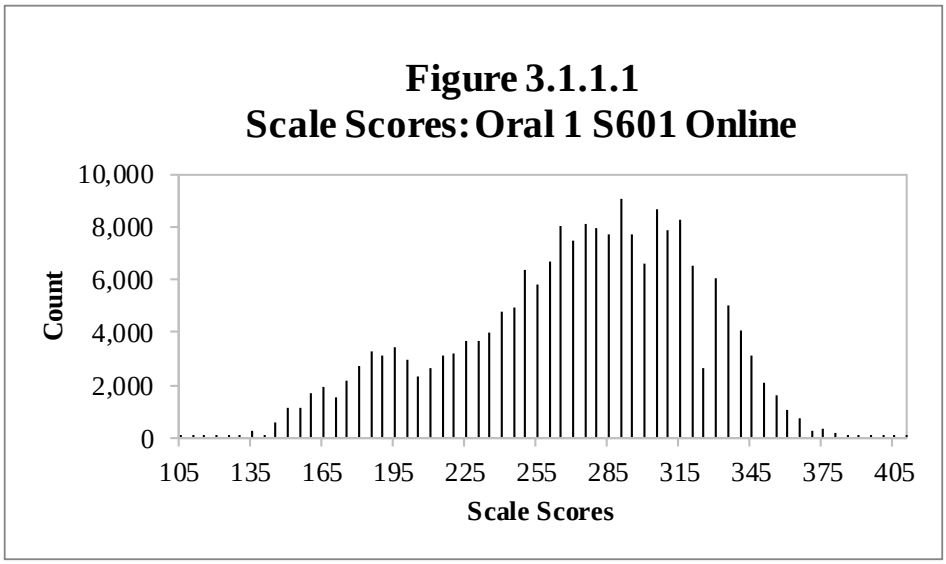
- The number of students in the analyses (count)
- The minimum observed scale score
- The maximum observed scale score
- The mean (average) scale score
- The standard deviation (std. dev.) of the scale score

3.1.1 Oral

3.1.1.1 Grade 1

Table 3.1.1.1
Scale Score Descriptive Statistics: Oral 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	198,705	109	411	271.81	50.41
Total	198,705	109	411	271.81	50.41

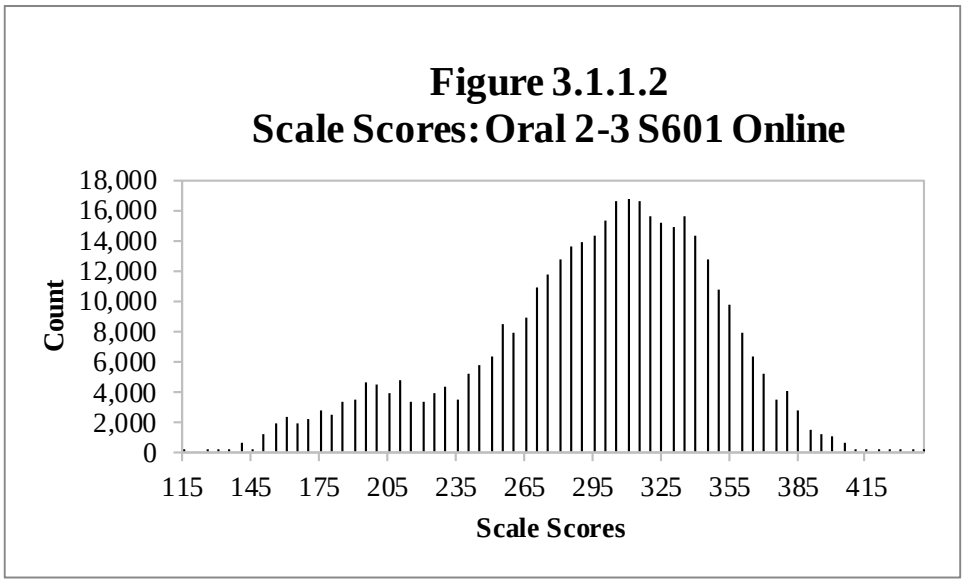


3.1.1.2 Grade 2-3

Table 3.1.1.2

Scale Score Descriptive Statistics: Oral 2-3 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	193,489	115	443	287.23	50.72
3	194,511	115	443	305.31	54.94
Total	388,000	115	443	296.30	53.65

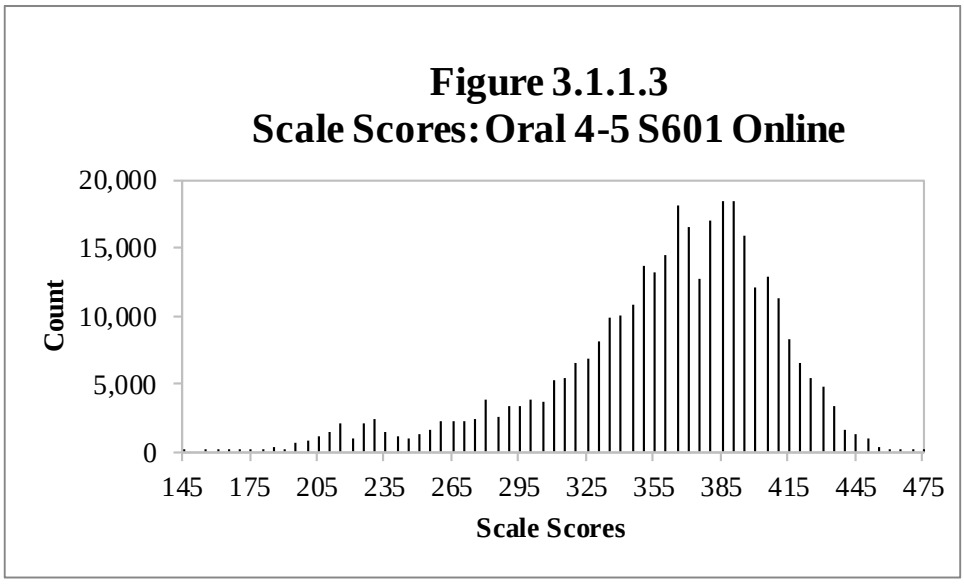


3.1.1.3 Grades 4-5

Table 3.1.1.3

Scale Score Descriptive Statistics: Oral 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	186,045	156	478	359.05	48.96
5	154,950	149	478	360.85	53.33
Total	340,995	149	478	359.87	51.00

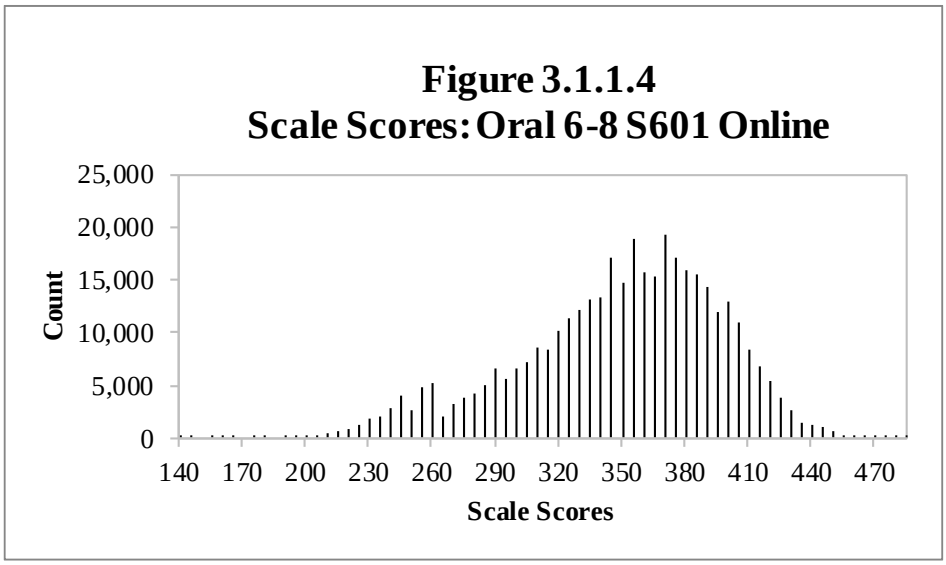


3.1.1.4 Grades 6-8

Table 3.1.1.4

Scale Score Descriptive Statistics: Oral 6-8 S601 Online

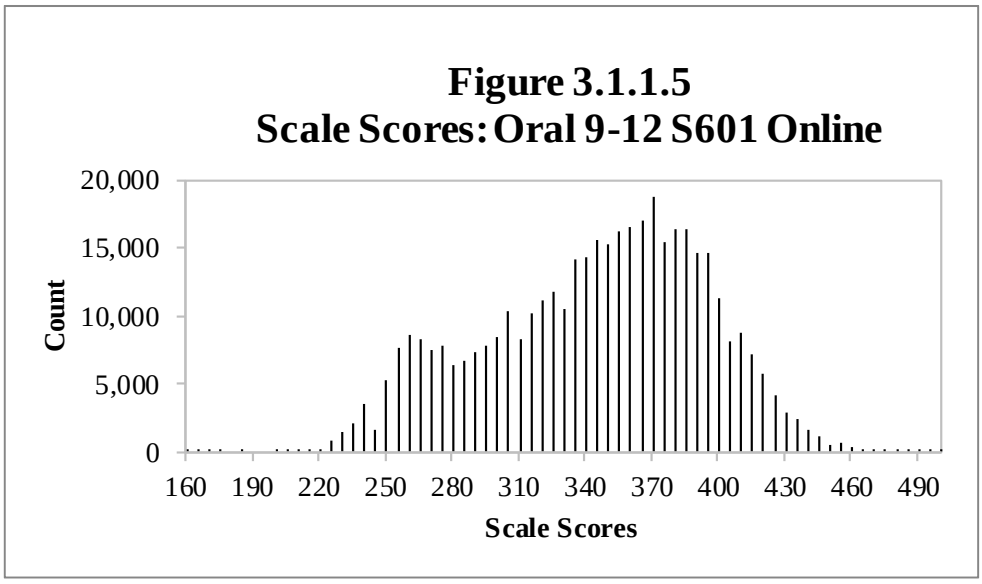
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	128,699	140	474	347.97	43.98
7	128,759	140	488	351.25	47.66
8	123,435	140	488	353.18	51.34
Total	380,893	140	488	350.77	47.75



3.1.1.5 Grades 9-12

Table 3.1.1.5
Scale Score Descriptive Statistics: Oral 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	144,063	160	501	339.55	49.27
10	117,258	166	491	345.48	49.56
11	84,028	160	501	352.25	49.65
12	70,302	172	501	352.73	49.46
Total	415,651	160	501	346.02	49.76

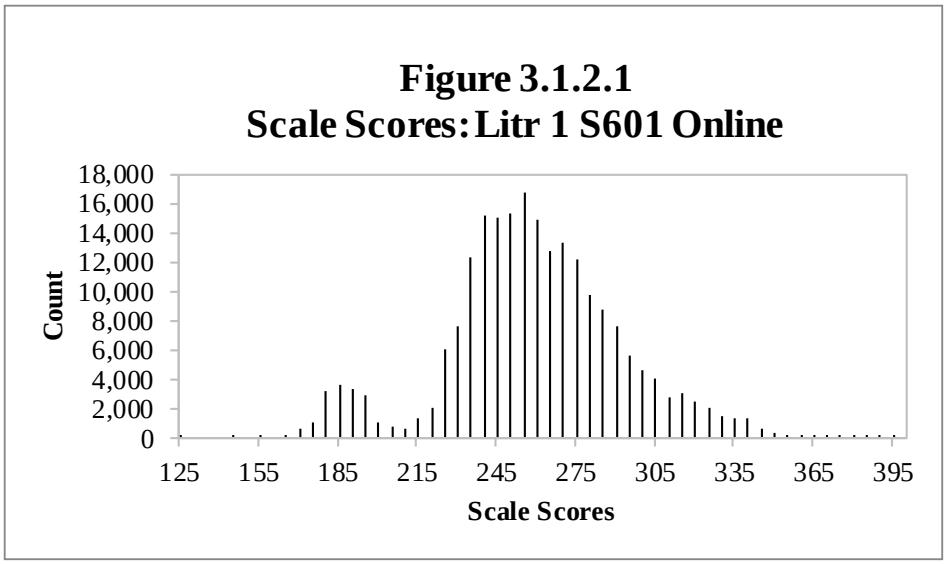


3.1.2 Literacy

3.1.2.1 Grade 1

Table 3.1.2.1
Scale Score Descriptive Statistics: Litr 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	218,829	126	424	260.90	33.52
Total	218,829	126	424	260.90	33.52

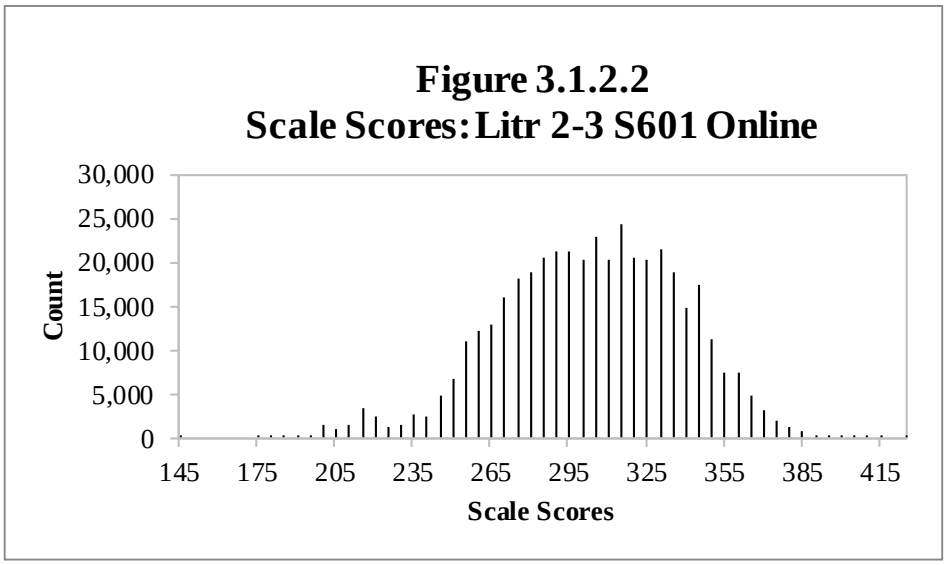


3.1.2.2 Grade 2-3

Table 3.1.2.2

Scale Score Descriptive Statistics: Litr 2-3 S601 Online

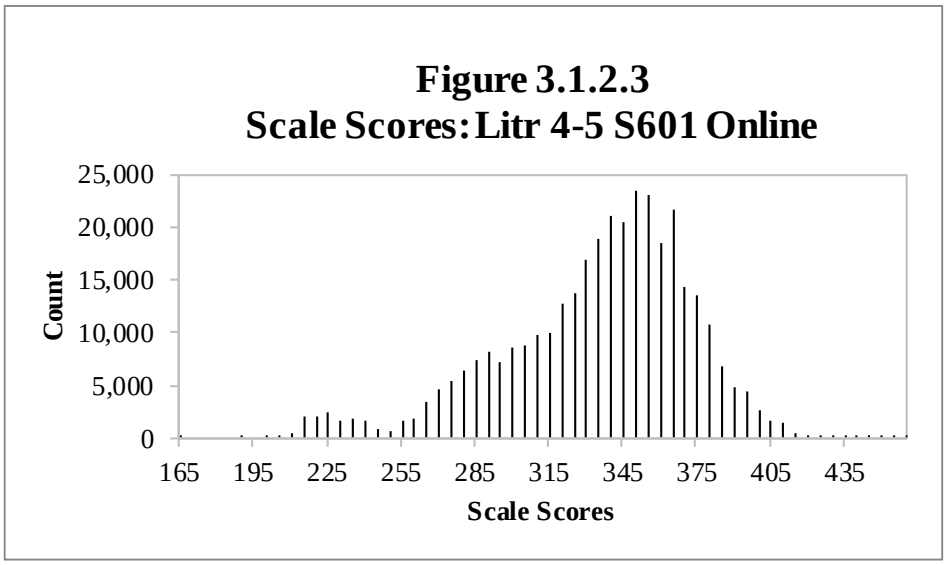
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	213,228	146	413	296.75	32.34
3	209,939	176	427	312.77	35.92
Total	423,167	146	427	304.70	35.09



3.1.2.3 Grades 4-5

Table 3.1.2.3
Scale Score Descriptive Statistics: Litr 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	190,734	165	453	333.65	38.19
5	159,114	194	461	339.38	39.62
Total	349,848	165	461	336.26	38.95

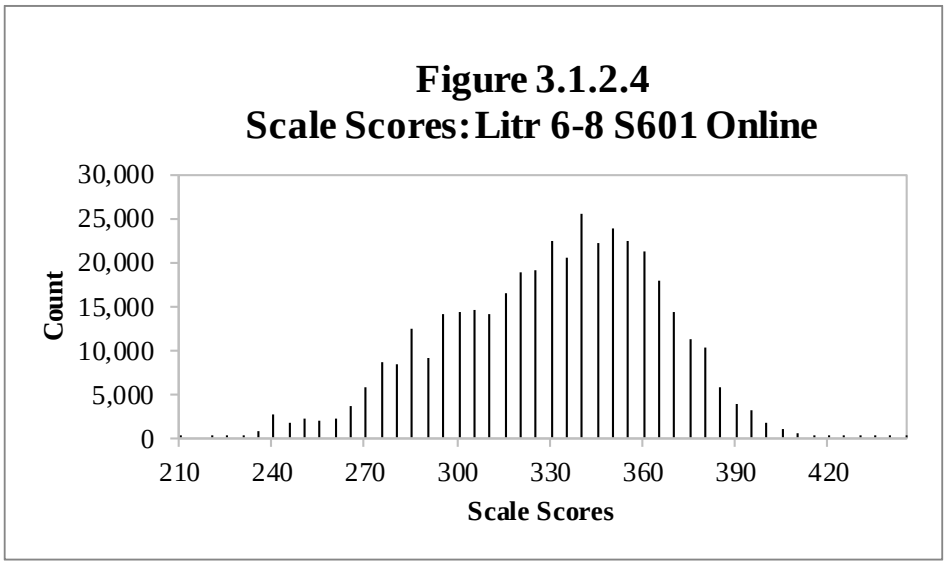


3.1.2.4 Grades 6-8

Table 3.1.2.4

Scale Score Descriptive Statistics: Litr 6-8 S601 Online

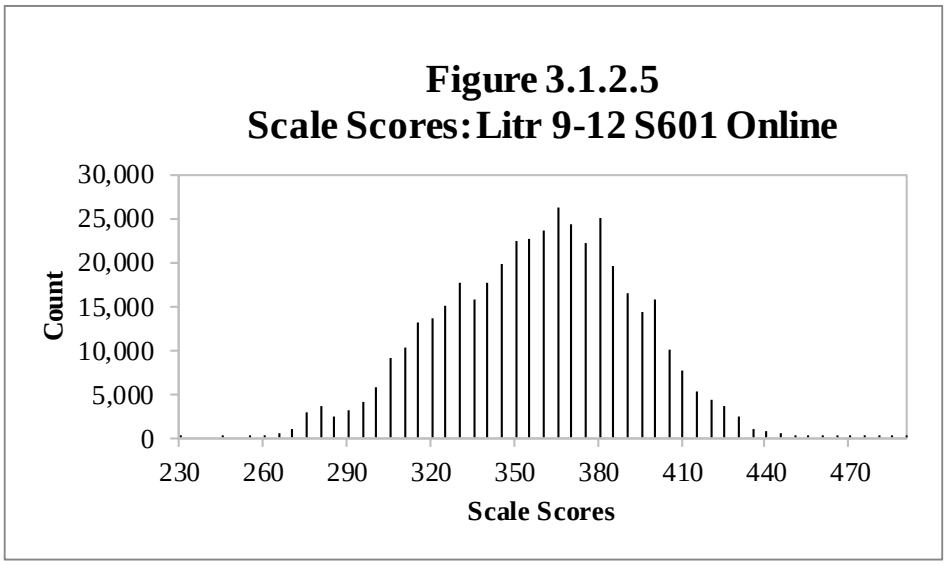
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	135,779	222	447	326.23	31.38
7	135,945	212	445	333.18	34.03
8	128,761	212	449	337.94	36.03
Total	400,485	212	449	332.36	34.17



3.1.2.5 Grades 9-12

Table 3.1.2.5
Scale Score Descriptive Statistics: Litr 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	147,803	246	497	354.13	34.52
10	119,687	246	477	359.06	33.56
11	86,316	246	486	365.28	33.19
12	71,178	233	492	365.95	32.50
Total	424,984	233	497	359.76	34.00

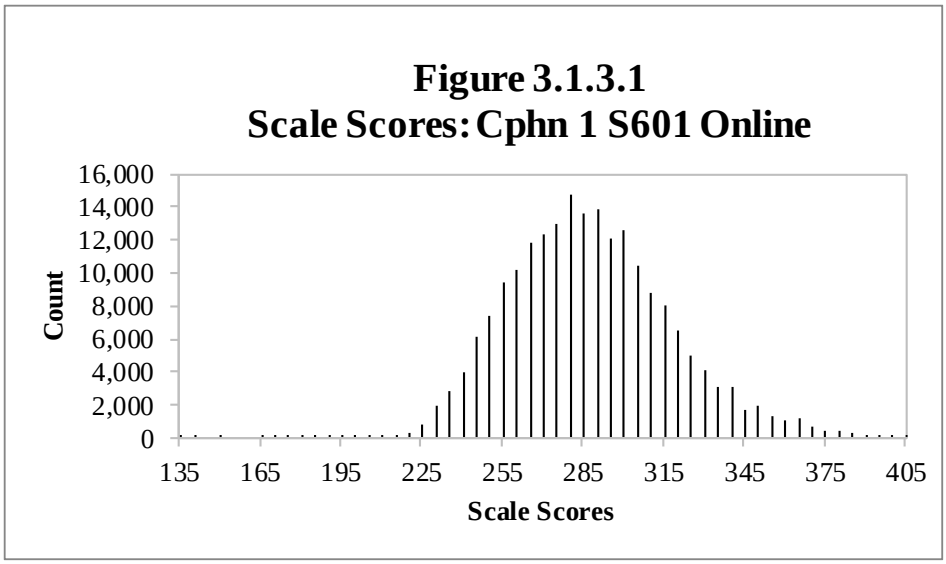


3.1.3 Comprehension

3.1.3.1 Grade 1

Table 3.1.3.1
Scale Score Descriptive Statistics: Cphn 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	206,299	135	407	289.56	30.07
Total	206,299	135	407	289.56	30.07

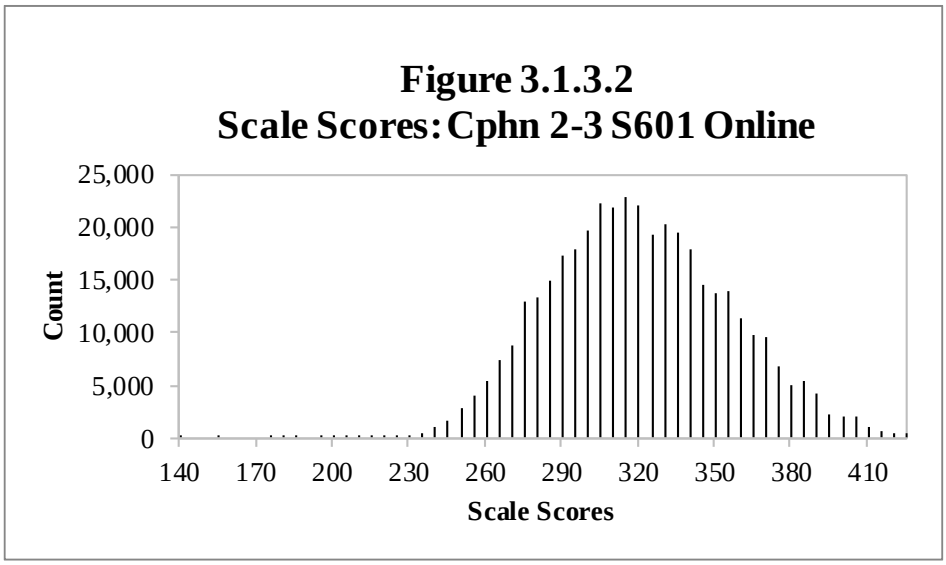


3.1.3.2 Grade 2-3

Table 3.1.3.2

Scale Score Descriptive Statistics: Cphn 2-3 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	199,328	144	437	315.36	30.89
3	198,479	179	437	329.37	37.61
Total	397,807	144	437	322.35	35.11

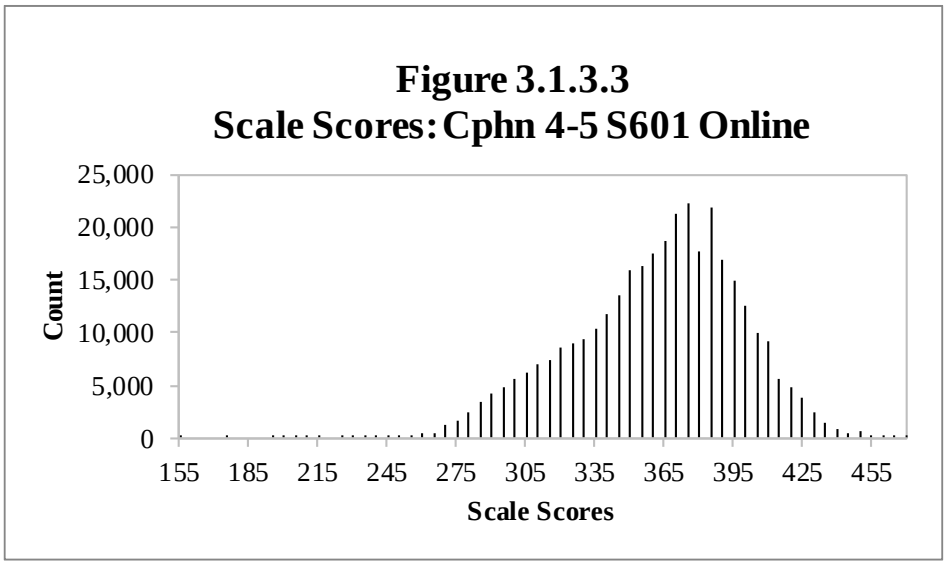


3.1.3.3 Grades 4-5

Table 3.1.3.3

Scale Score Descriptive Statistics: Cphn 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	187,543	177	473	362.28	34.65
5	156,288	159	473	365.84	37.51
Total	343,831	159	473	363.90	36.02

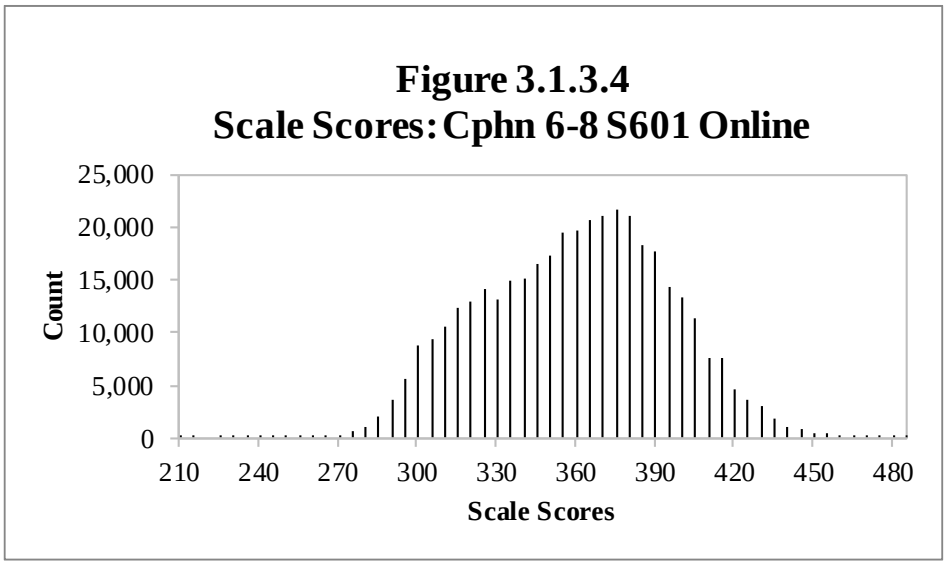


3.1.3.4 Grades 6-8

Table 3.1.3.4

Scale Score Descriptive Statistics: Cphn 6-8 S601 Online

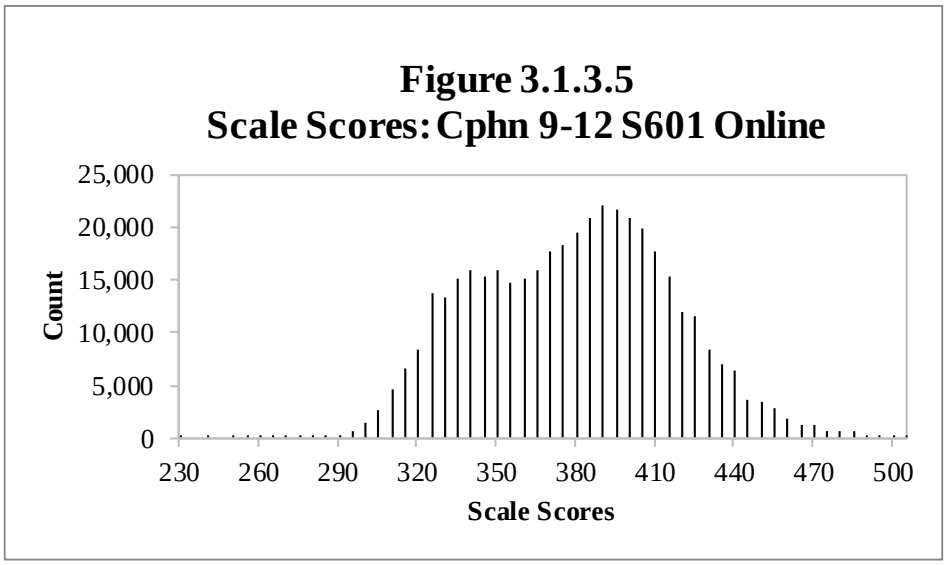
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	131,605	234	485	356.46	31.65
7	132,021	234	481	362.38	35.15
8	125,964	213	485	366.32	38.19
Total	389,590	213	485	361.66	35.28



3.1.3.5 Grades 9-12

Table 3.1.3.5
Scale Score Descriptive Statistics: Cphn 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	144,409	240	505	375.10	36.08
10	117,455	256	505	380.37	37.29
11	84,990	250	505	386.40	37.75
12	70,042	232	505	387.36	37.48
Total	416,896	232	505	380.95	37.34

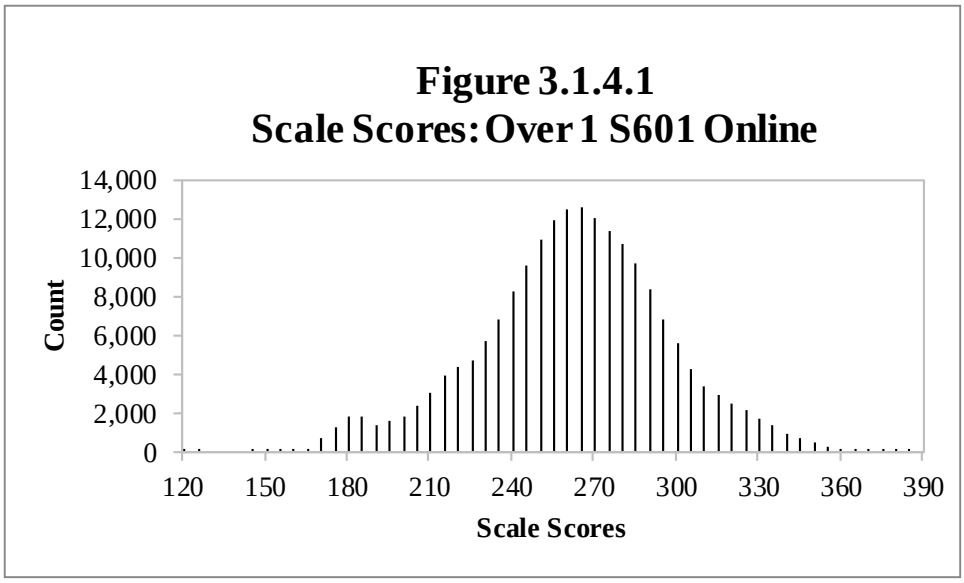


3.1.4 Overall

3.1.4.1 Grade 1

Table 3.1.4.1
Scale Score Descriptive Statistics: Over 1 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	192,851	122	409	263.97	34.49
Total	192,851	122	409	263.97	34.49

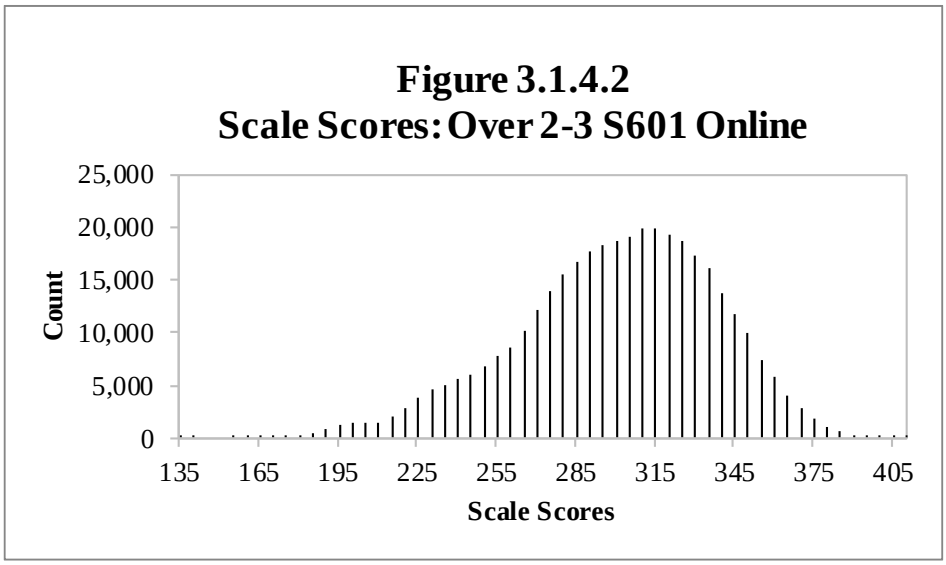


3.1.4.2 Grade 2-3

Table 3.1.4.2

Scale Score Descriptive Statistics: Over 2-3 S601 Online

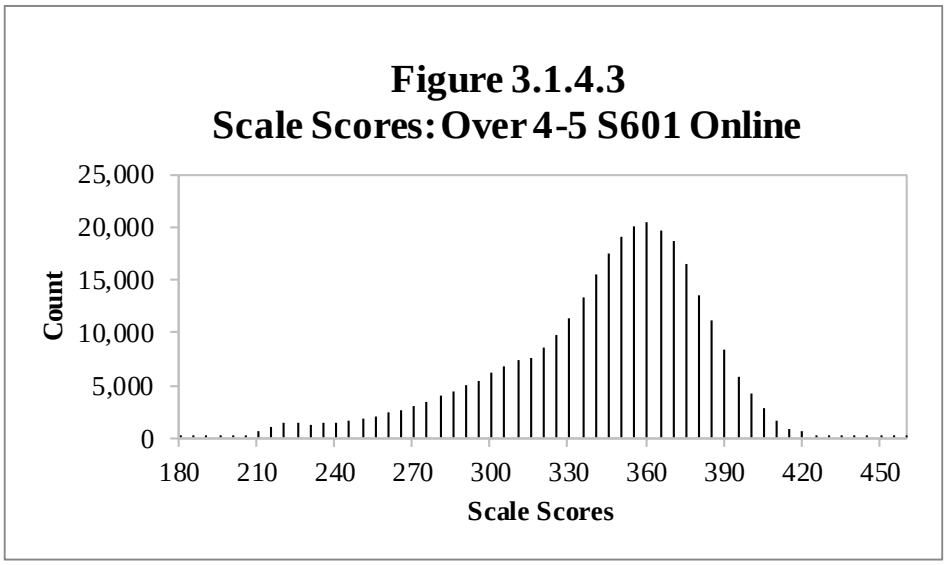
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	186,234	136	402	293.69	34.73
3	187,098	172	422	310.21	38.87
Total	373,332	136	422	301.97	37.77



3.1.4.3 Grades 4-5

Table 3.1.4.3
Scale Score Descriptive Statistics: Over 4-5 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	170,696	184	446	341.25	38.98
5	142,983	192	460	345.62	41.41
Total	313,679	184	460	343.24	40.16

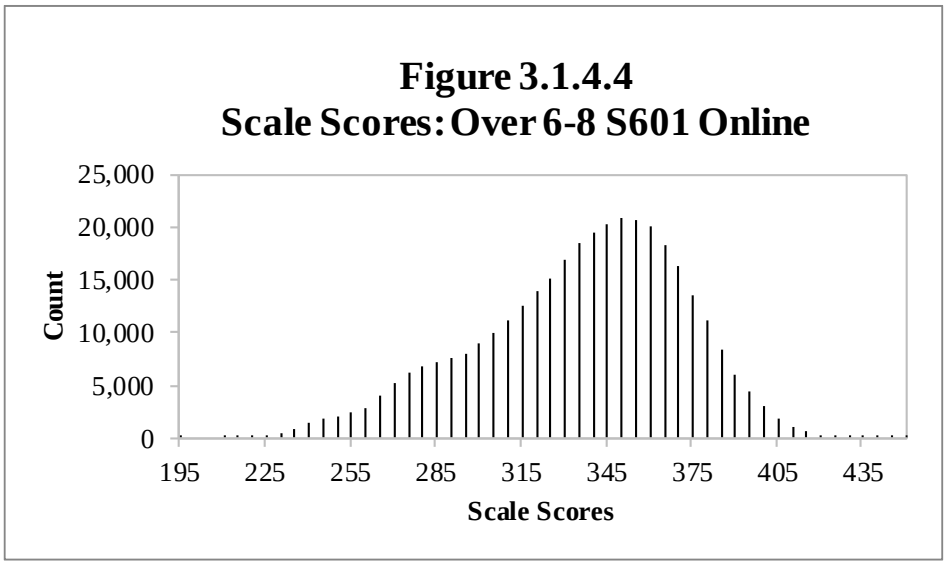


3.1.4.4 Grades 6-8

Table 3.1.4.4

Scale Score Descriptive Statistics: Over 6-8 S601 Online

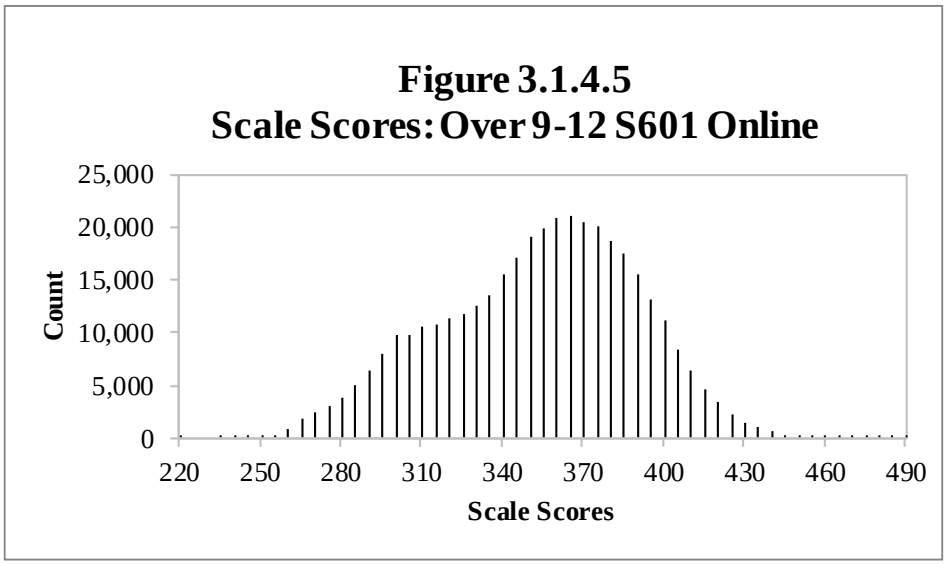
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	118,160	218	444	332.63	32.85
7	118,892	218	448	338.46	36.01
8	113,911	197	454	342.10	38.58
Total	350,963	197	454	337.68	36.07



3.1.4.5 Grades 9-12

Table 3.1.4.5
Scale Score Descriptive Statistics: Over 9-12 S601 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	131,765	239	493	349.39	37.03
10	107,284	236	480	354.60	36.40
11	77,102	243	487	360.88	36.14
12	64,881	223	476	361.58	35.33
Total	381,032	223	493	355.26	36.73



3.2 Proficiency Level Distribution for Composites

Figures and tables in this section provide information on the proficiency level distribution for each of the composites for each grade-level cluster.

In each figure, the horizontal axis shows the six WIDA proficiency levels. The vertical axis shows the percentage of students. Each bar shows the percentage of students who were placed into each proficiency level in the domain being tested on this test form.

The tables in this section present, by grade and by total for the grade-level cluster:

- The WIDA proficiency level designation (1–6)
- The number of students (count) whose performance on the test form placed them into that proficiency level in the domain being tested
- The percentage of students, out of the total number of students taking the form, who were placed into that proficiency level in the domain being tested

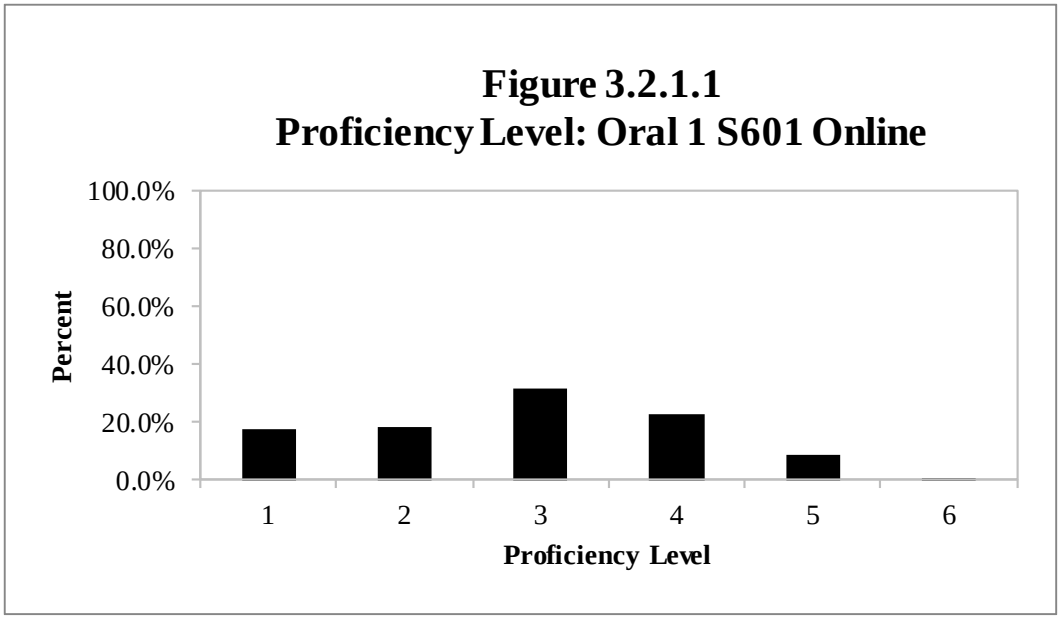
3.2.1 Oral

3.2.1.1 Grade 1

Table 3.2.1.1

Proficiency Level Distribution: Oral 1 S601 Online

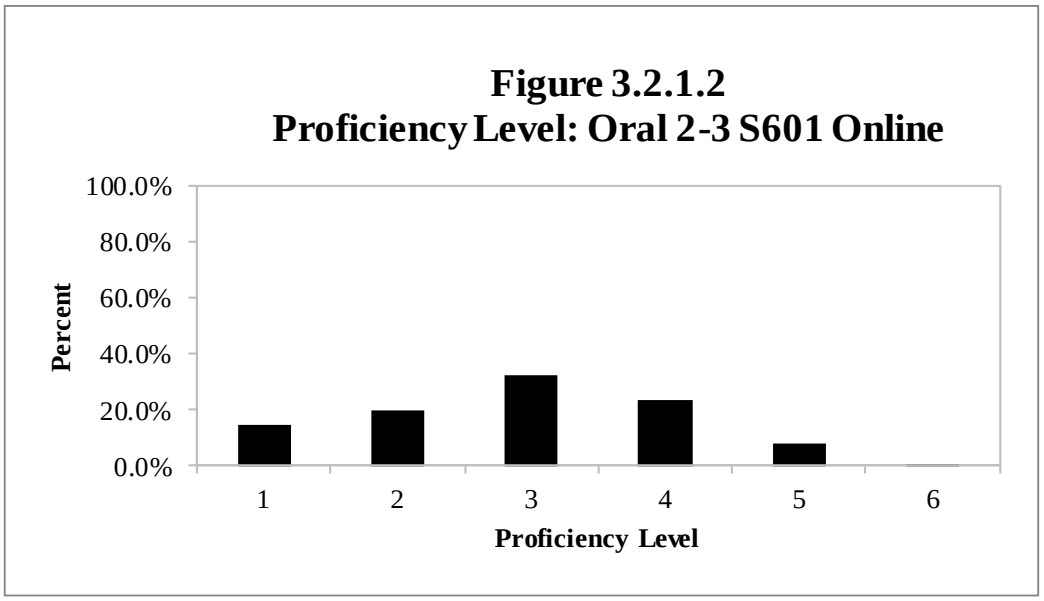
	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	34,866	17.55%	34,866	17.55%
2	35,930	18.08%	35,930	18.08%
3	63,538	31.98%	63,538	31.98%
4	45,261	22.78%	45,261	22.78%
5	17,389	8.75%	17,389	8.75%
6	1,721	0.87%	1,721	0.87%
Total	198,705	100.00%	198,705	100.00%



3.2.1.2 Grade 2-3

Table 3.2.1.2
Proficiency Level Distribution: Oral 2-3 S601 Online

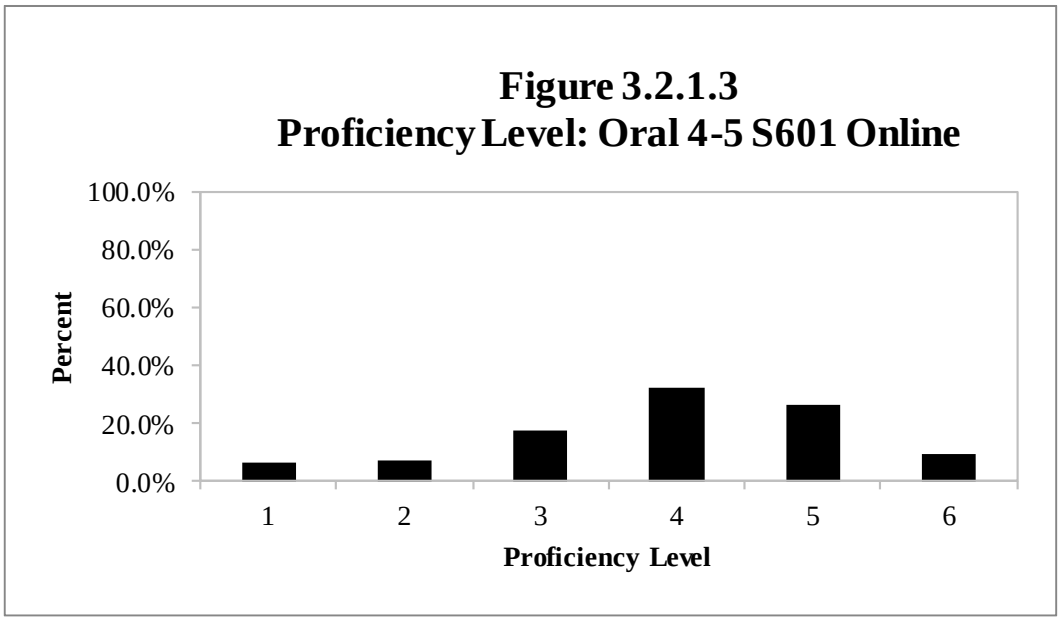
	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	28,686	14.83%	28,784	14.80%	57,470	14.81%
2	43,442	22.45%	33,328	17.13%	76,770	19.79%
3	62,748	32.43%	63,917	32.86%	126,665	32.65%
4	43,086	22.27%	49,443	25.42%	92,529	23.85%
5	13,912	7.19%	17,047	8.76%	30,959	7.98%
6	1,615	0.83%	1,992	1.02%	3,607	0.93%
Total	193,489	100.00%	194,511	100.00%	388,000	100.00%



3.2.1.3 Grades 4-5

Table 3.2.1.3
Proficiency Level Distribution: Oral 4-5 S601 Online

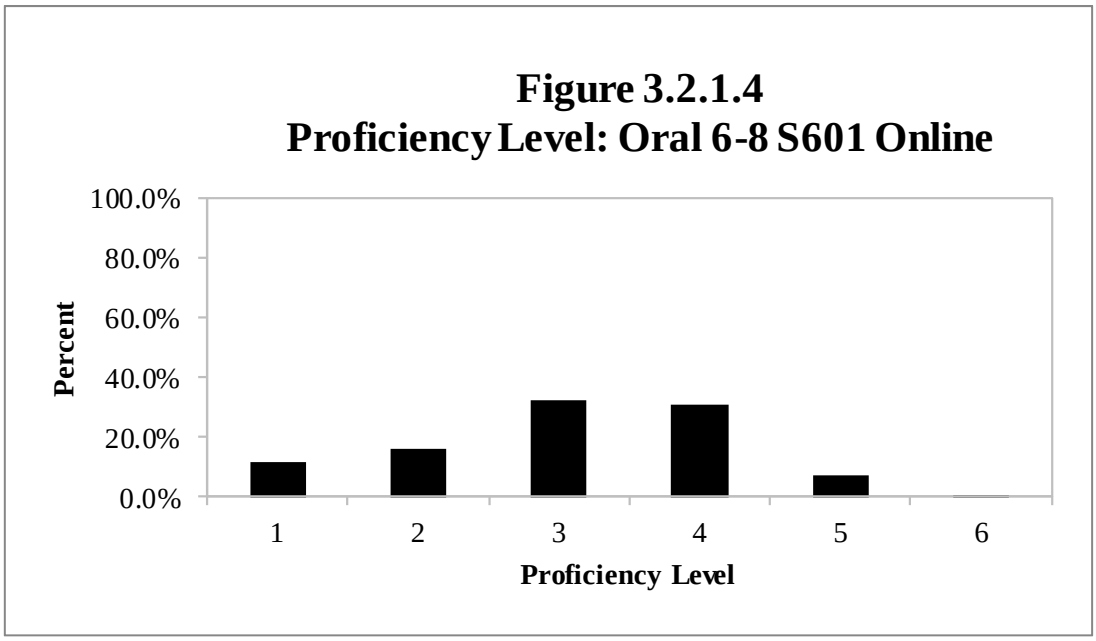
	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	9,541	5.13%	12,350	7.97%	21,891	6.42%
2	13,765	7.40%	12,148	7.84%	25,913	7.60%
3	32,023	17.21%	27,470	17.73%	59,493	17.45%
4	58,687	31.54%	51,917	33.51%	110,604	32.44%
5	51,954	27.93%	39,602	25.56%	91,556	26.85%
6	20,075	10.79%	11,463	7.40%	31,538	9.25%
Total	186,045	100.00%	154,950	100.00%	340,995	100.00%



3.2.1.4 Grades 6-8

Table 3.2.1.4
Proficiency Level Distribution: Oral 6-8 S601 Online

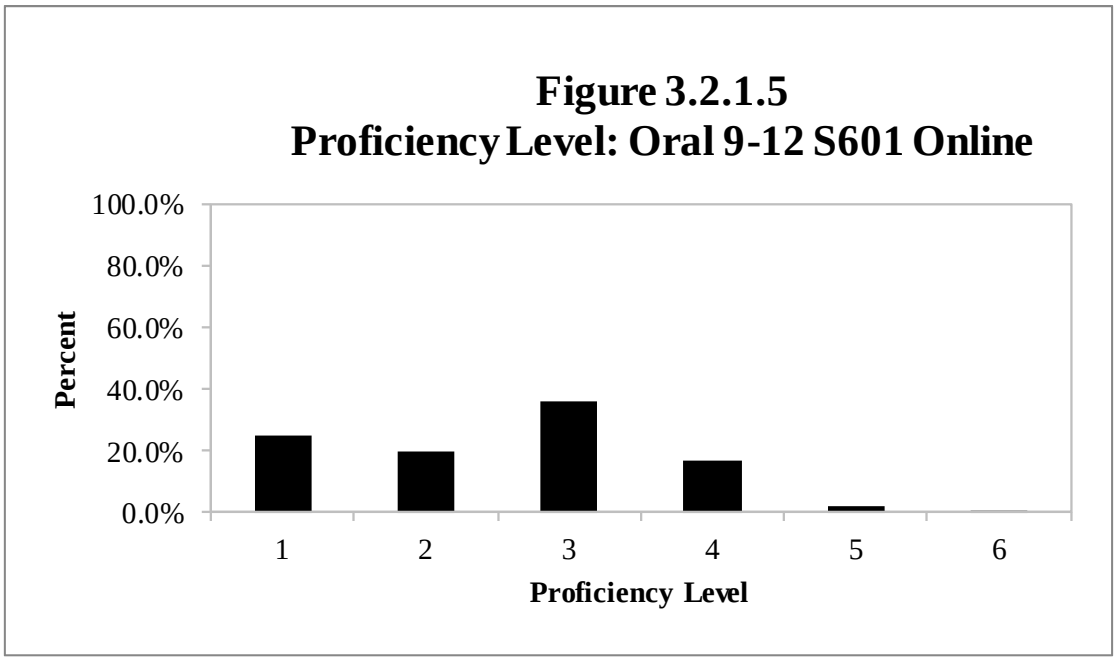
	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	10,789	8.38%	14,877	11.55%	18,427	14.93%	44,093	11.58%
2	19,791	15.38%	21,205	16.47%	20,056	16.25%	61,052	16.03%
3	43,228	33.59%	41,070	31.90%	39,696	32.16%	123,994	32.55%
4	41,939	32.59%	40,516	31.47%	36,184	29.31%	118,639	31.15%
5	11,674	9.07%	9,660	7.50%	7,697	6.24%	29,031	7.62%
6	1,278	0.99%	1,431	1.11%	1,375	1.11%	4,084	1.07%
Total	128,699	100.00%	128,759	100.00%	123,435	100.00%	380,893	100.00%



3.2.1.5 Grades 9-12

Table 3.2.1.5
Proficiency Level Distribution: Oral 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	35,731	24.80%	29,559	25.21%	19,965	23.76%	17,859	25.40%	103,114	24.81%
2	29,881	20.74%	23,367	19.93%	16,566	19.71%	13,895	19.76%	83,709	20.14%
3	49,714	34.51%	41,362	35.27%	31,183	37.11%	27,518	39.14%	149,777	36.03%
4	25,635	17.79%	20,361	17.36%	14,253	16.96%	9,849	14.01%	70,098	16.86%
5	2,763	1.92%	2,281	1.95%	1,811	2.16%	1,036	1.47%	7,891	1.90%
6	339	0.24%	328	0.28%	250	0.30%	145	0.21%	1,062	0.26%
Total	144,063	100.00%	117,258	100.00%	84,028	100.00%	70,302	100.00%	415,651	100.00%



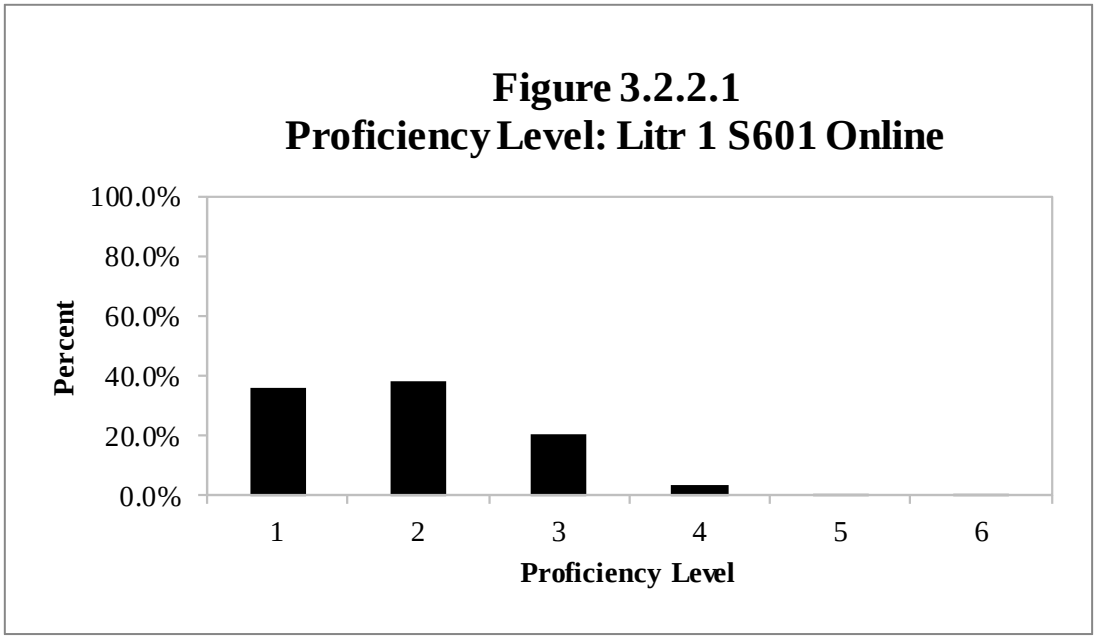
3.2.2 Literacy

3.2.2.1 Grade 1

Table 3.2.2.1

Proficiency Level Distribution: Litr 1 S601 Online

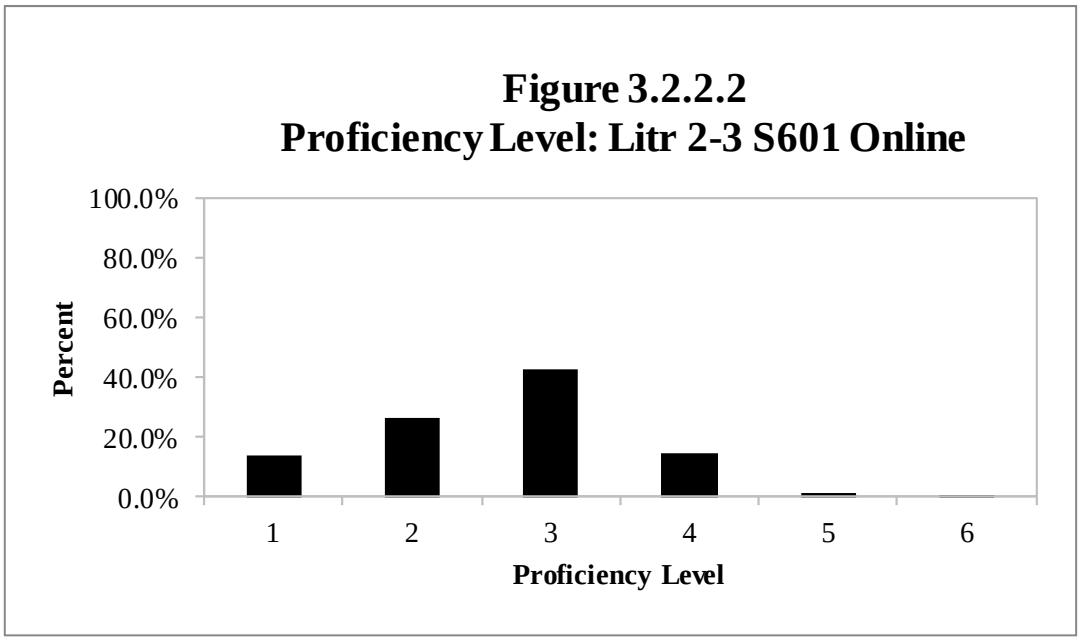
	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	79,072	36.13%	79,072	36.13%
2	84,332	38.54%	84,332	38.54%
3	45,354	20.73%	45,354	20.73%
4	8,427	3.85%	8,427	3.85%
5	1,524	0.70%	1,524	0.70%
6	120	0.05%	120	0.05%
Total	218,829	100.00%	218,829	100.00%



3.2.2.2 Grade 2-3

Table 3.2.2.2
Proficiency Level Distribution: Litr 2-3 S601 Online

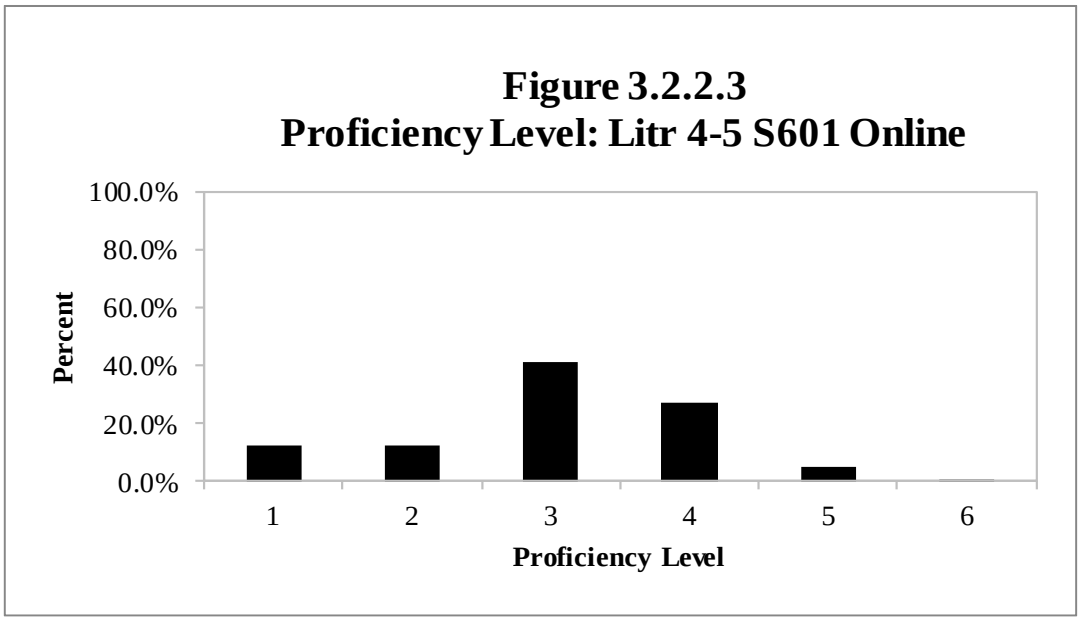
	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	28,659	13.44%	29,166	13.89%	57,825	13.66%
2	65,067	30.52%	47,478	22.62%	112,545	26.60%
3	92,978	43.60%	89,499	42.63%	182,477	43.12%
4	24,242	11.37%	38,962	18.56%	63,204	14.94%
5	2,100	0.98%	4,521	2.15%	6,621	1.56%
6	182	0.09%	313	0.15%	495	0.12%
Total	213,228	100.00%	209,939	100.00%	423,167	100.00%



3.2.2.3 Grades 4-5

Table 3.2.2.3
Proficiency Level Distribution: Litr 4-5 S601 Online

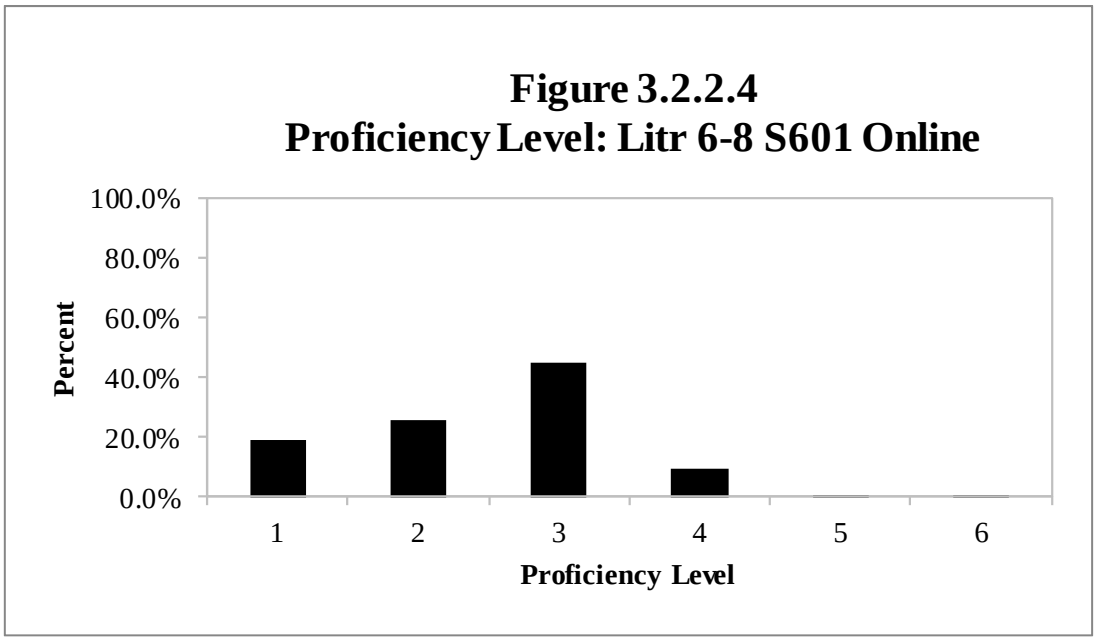
	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	22,540	11.82%	20,959	13.17%	43,499	12.43%
2	23,284	12.21%	19,722	12.39%	43,006	12.29%
3	81,071	42.50%	64,231	40.37%	145,302	41.53%
4	51,986	27.26%	44,714	28.10%	96,700	27.64%
5	9,761	5.12%	8,219	5.17%	17,980	5.14%
6	2,092	1.10%	1,269	0.80%	3,361	0.96%
Total	190,734	100.00%	159,114	100.00%	349,848	100.00%



3.2.2.4 Grades 6-8

Table 3.2.2.4
Proficiency Level Distribution: Litr 6-8 S601 Online

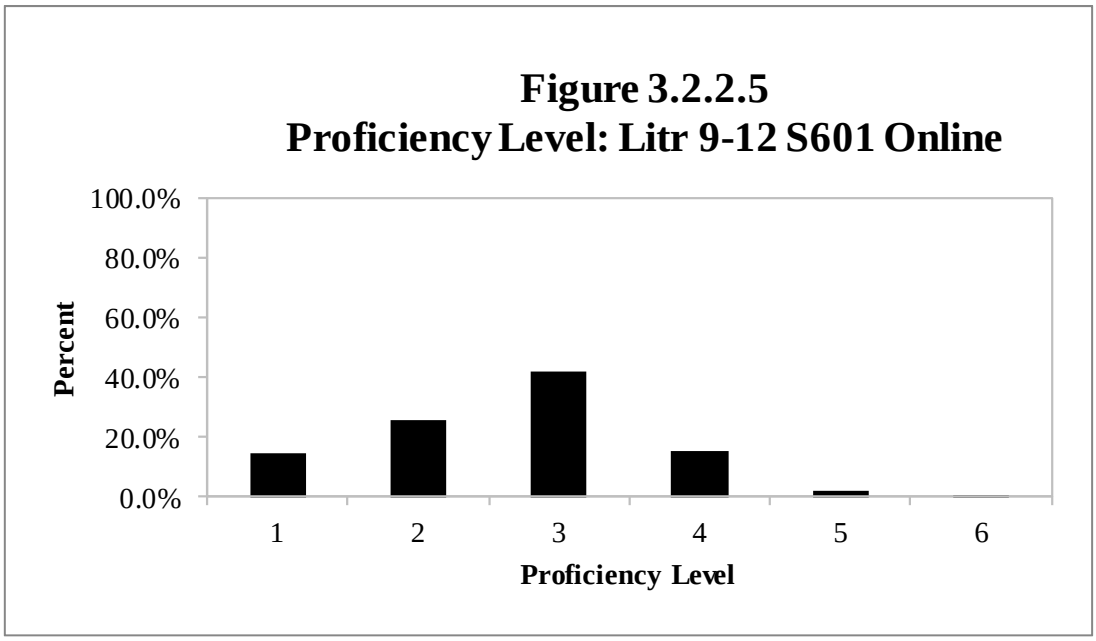
	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	24,615	18.13%	25,625	18.85%	27,649	21.47%	77,889	19.45%
2	35,880	26.43%	35,093	25.81%	31,722	24.64%	102,695	25.64%
3	64,614	47.59%	61,511	45.25%	54,239	42.12%	180,364	45.04%
4	10,206	7.52%	12,972	9.54%	14,427	11.20%	37,605	9.39%
5	427	0.31%	726	0.53%	708	0.55%	1,861	0.46%
6	37	0.03%	18	0.01%	16	0.01%	71	0.02%
Total	135,779	100.00%	135,945	100.00%	128,761	100.00%	400,485	100.00%



3.2.2.5 Grades 9-12

Table 3.2.2.5
Proficiency Level Distribution: Litr 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	21,406	14.48%	16,815	14.05%	12,039	13.95%	12,570	17.66%	62,830	14.78%
2	34,861	23.59%	30,759	25.70%	22,805	26.42%	20,759	29.16%	109,184	25.69%
3	63,047	42.66%	49,804	41.61%	35,814	41.49%	28,729	40.36%	177,394	41.74%
4	24,648	16.68%	19,295	16.12%	13,357	15.47%	7,941	11.16%	65,241	15.35%
5	3,605	2.44%	2,887	2.41%	2,238	2.59%	1,169	1.64%	9,899	2.33%
6	236	0.16%	127	0.11%	63	0.07%	10	0.01%	436	0.10%
Total	147,803	100.00%	119,687	100.00%	86,316	100.00%	71,178	100.00%	424,984	100.00%



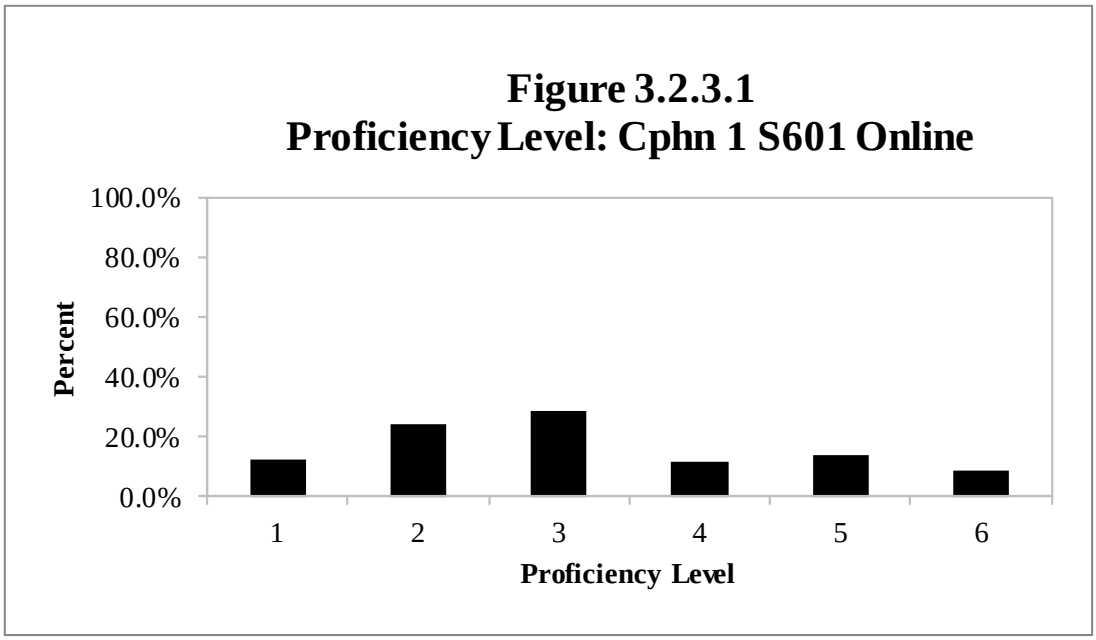
3.2.3 Comprehension

3.2.3.1 Grade 1

Table 3.2.3.1

Proficiency Level Distribution: Cphn 1 S601 Online

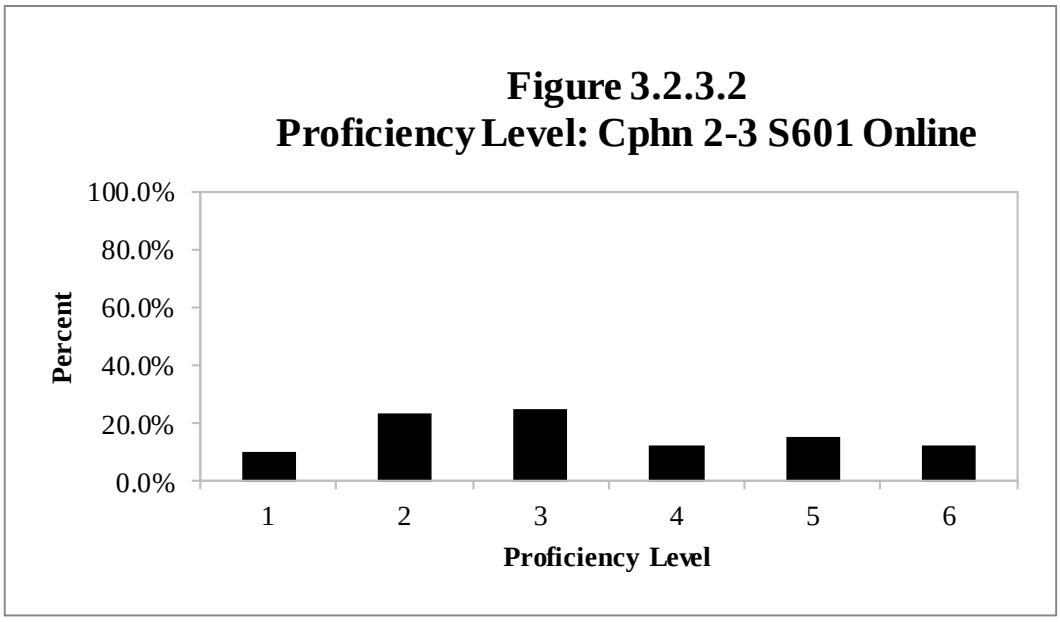
	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	25,296	12.26%	25,296	12.26%
2	49,808	24.14%	49,808	24.14%
3	59,706	28.94%	59,706	28.94%
4	24,854	12.05%	24,854	12.05%
5	28,370	13.75%	28,370	13.75%
6	18,265	8.85%	18,265	8.85%
Total	206,299	100.00%	206,299	100.00%



3.2.3.2 Grade 2-3

Table 3.2.3.2
Proficiency Level Distribution: Cphn 2-3 S601 Online

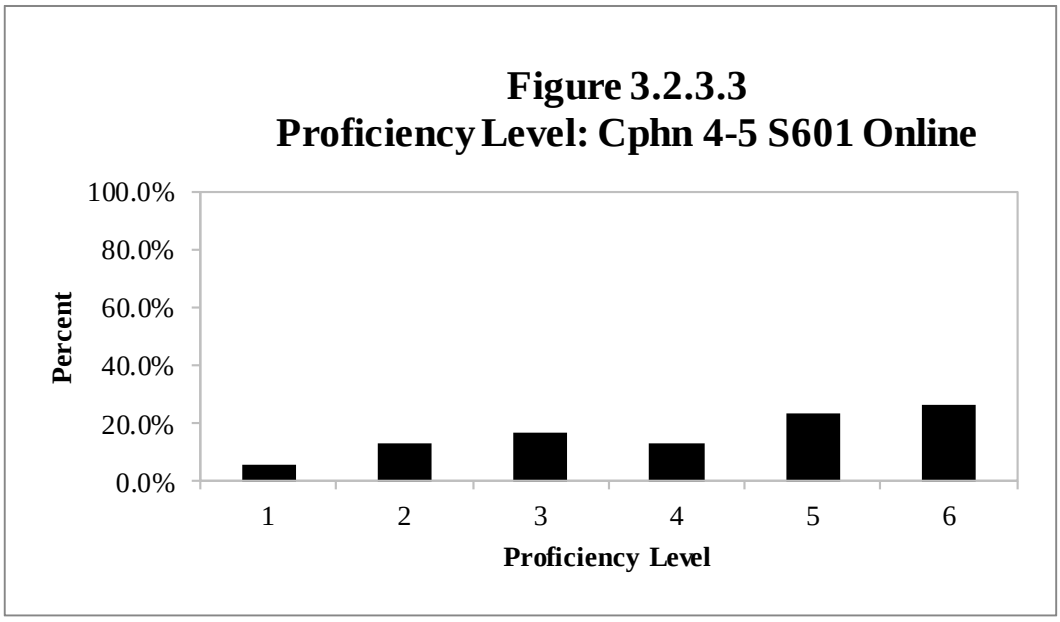
	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	13,878	6.96%	27,646	13.93%	41,524	10.44%
2	48,522	24.34%	45,651	23.00%	94,173	23.67%
3	56,540	28.37%	43,669	22.00%	100,209	25.19%
4	27,922	14.01%	21,704	10.94%	49,626	12.47%
5	30,778	15.44%	30,385	15.31%	61,163	15.38%
6	21,688	10.88%	29,424	14.82%	51,112	12.85%
Total	199,328	100.00%	198,479	100.00%	397,807	100.00%



3.2.3.3 Grades 4-5

Table 3.2.3.3
Proficiency Level Distribution: Cphn 4-5 S601 Online

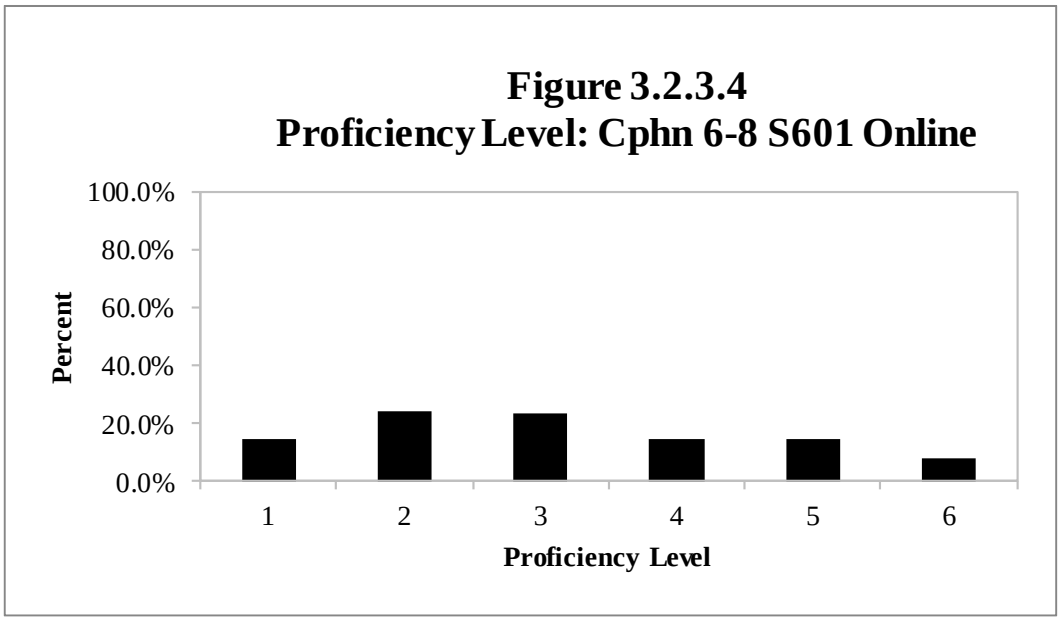
	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	7,487	3.99%	13,485	8.63%	20,972	6.10%
2	24,849	13.25%	21,328	13.65%	46,177	13.43%
3	31,749	16.93%	25,709	16.45%	57,458	16.71%
4	24,767	13.21%	21,793	13.94%	46,560	13.54%
5	46,278	24.68%	36,163	23.14%	82,441	23.98%
6	52,413	27.95%	37,810	24.19%	90,223	26.24%
Total	187,543	100.00%	156,288	100.00%	343,831	100.00%



3.2.3.4 Grades 6-8

Table 3.2.3.4
Proficiency Level Distribution: Cphn 6-8 S601 Online

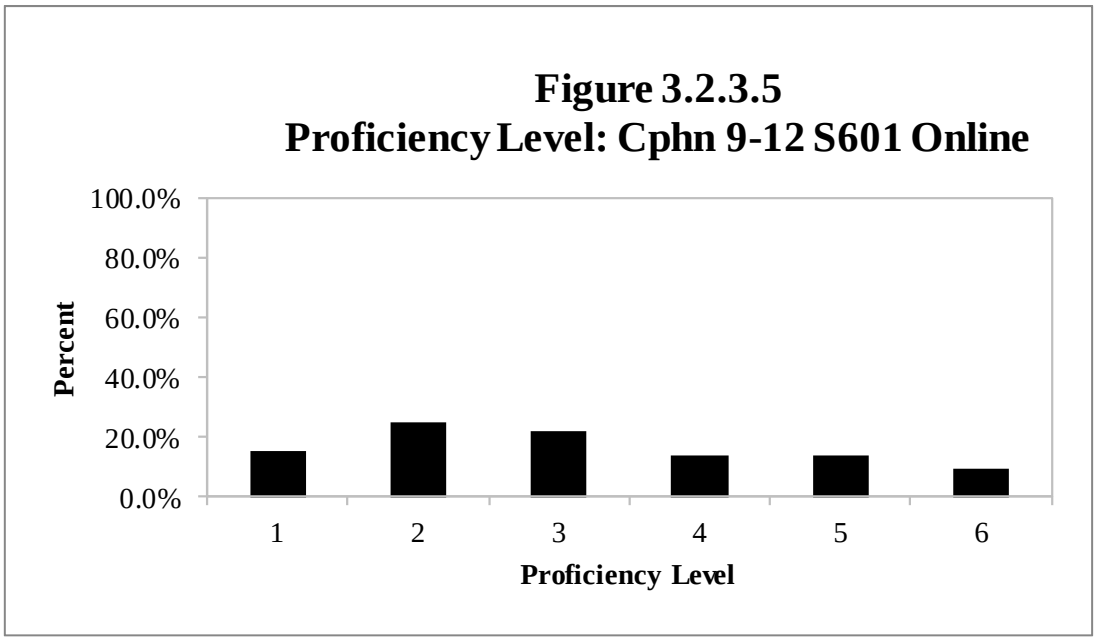
	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	14,219	10.80%	19,096	14.46%	22,964	18.23%	56,279	14.45%
2	34,079	25.89%	32,277	24.45%	29,501	23.42%	95,857	24.60%
3	35,251	26.79%	31,003	23.48%	26,466	21.01%	92,720	23.80%
4	20,195	15.35%	20,196	15.30%	17,409	13.82%	57,800	14.84%
5	19,358	14.71%	18,653	14.13%	18,647	14.80%	56,658	14.54%
6	8,503	6.46%	10,796	8.18%	10,977	8.71%	30,276	7.77%
Total	131,605	100.00%	132,021	100.00%	125,964	100.00%	389,590	100.00%



3.2.3.5 Grades 9-12

Table 3.2.3.5
Proficiency Level Distribution: Cphn 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	19,796	13.71%	18,090	15.40%	13,248	15.59%	12,400	17.70%	63,534	15.24%
2	37,464	25.94%	29,018	24.71%	20,289	23.87%	17,310	24.71%	104,081	24.97%
3	32,142	22.26%	26,126	22.24%	18,847	22.18%	15,995	22.84%	93,110	22.33%
4	20,813	14.41%	16,312	13.89%	11,091	13.05%	9,438	13.47%	57,654	13.83%
5	21,182	14.67%	15,925	13.56%	12,337	14.52%	8,603	12.28%	58,047	13.92%
6	13,012	9.01%	11,984	10.20%	9,178	10.80%	6,296	8.99%	40,470	9.71%
Total	144,409	100.00%	117,455	100.00%	84,990	100.00%	70,042	100.00%	416,896	100.00%



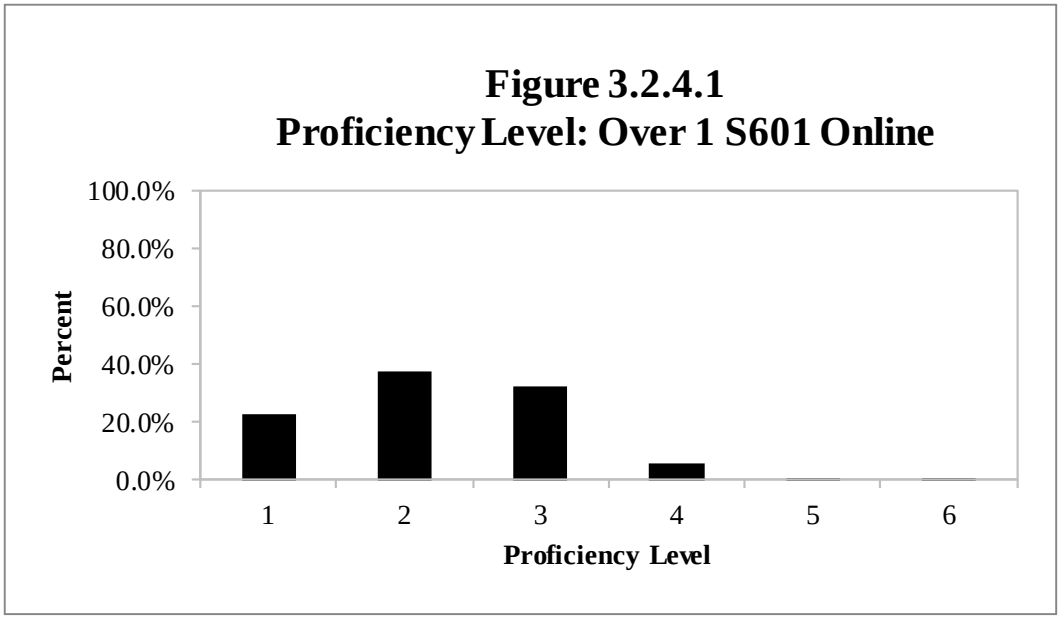
3.2.4 Overall

3.2.4.1 Grade 1

Table 3.2.4.1

Proficiency Level Distribution: Over 1 S601 Online

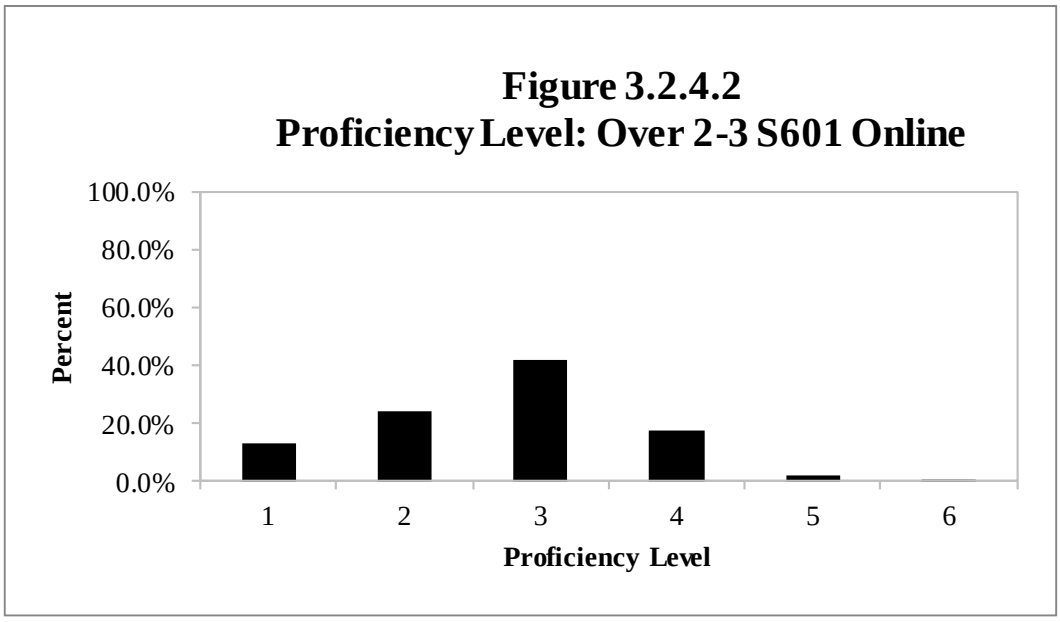
	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	44,722	23.19%	44,722	23.19%
2	72,324	37.50%	72,324	37.50%
3	62,516	32.42%	62,516	32.42%
4	11,386	5.90%	11,386	5.90%
5	1,814	0.94%	1,814	0.94%
6	89	0.05%	89	0.05%
Total	192,851	100.00%	192,851	100.00%



3.2.4.2 Grade 2-3

Table 3.2.4.2
Proficiency Level Distribution: Over 2-3 S601 Online

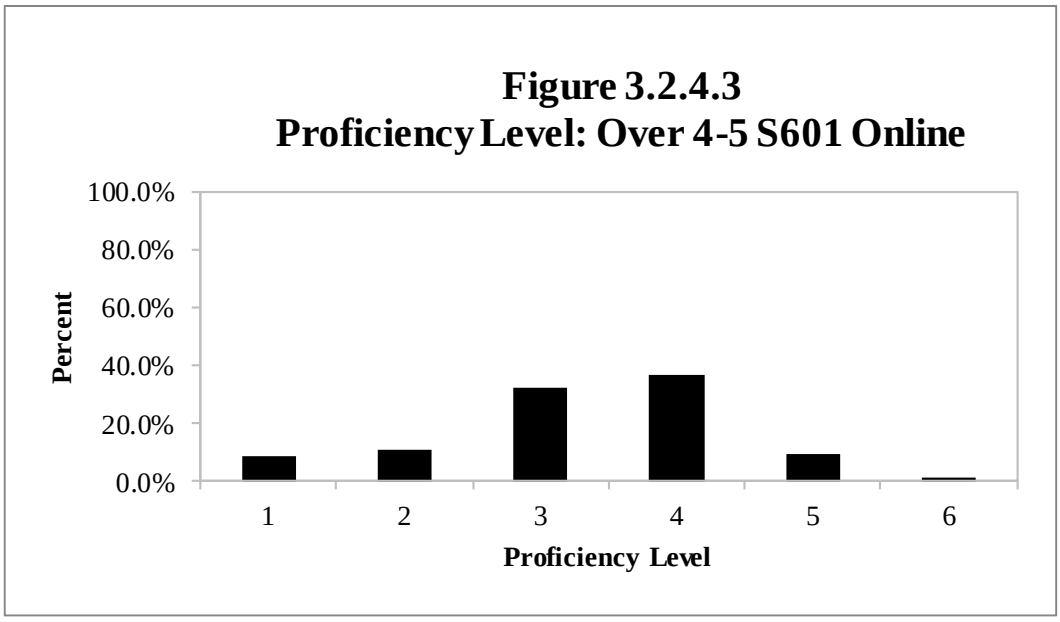
	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	24,373	13.09%	25,094	13.41%	49,467	13.25%
2	52,304	28.09%	39,726	21.23%	92,030	24.65%
3	80,456	43.20%	77,190	41.26%	157,646	42.23%
4	26,229	14.08%	39,690	21.21%	65,919	17.66%
5	2,721	1.46%	5,227	2.79%	7,948	2.13%
6	151	0.08%	171	0.09%	322	0.09%
Total	186,234	100.00%	187,098	100.00%	373,332	100.00%



3.2.4.3 Grades 4-5

Table 3.2.4.3
Proficiency Level Distribution: Over 4-5 S601 Online

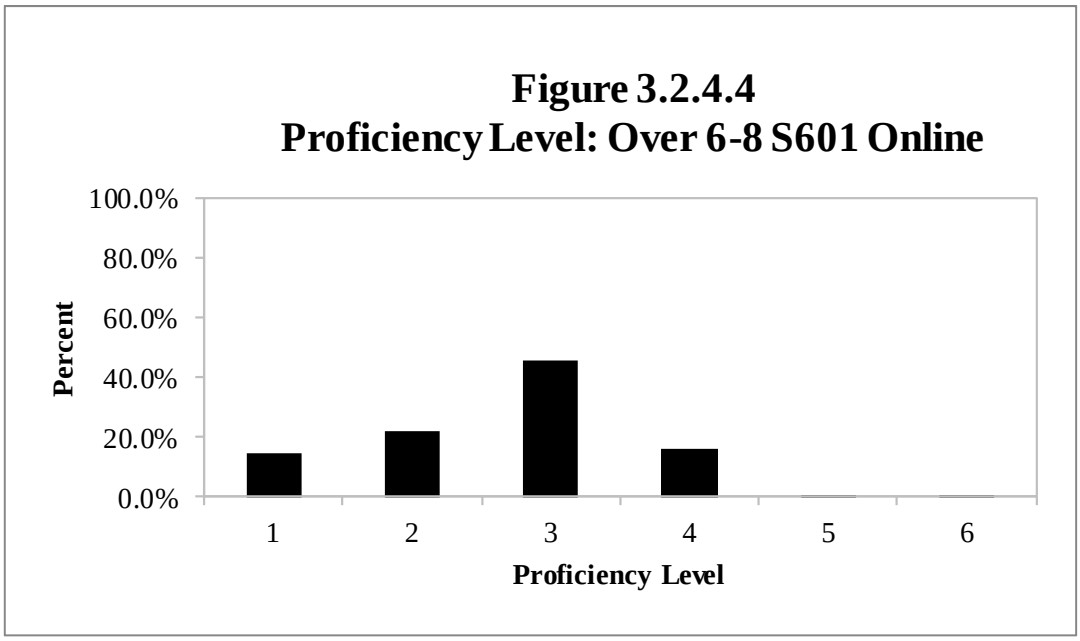
	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	13,386	7.84%	14,500	10.14%	27,886	8.89%
2	18,042	10.57%	15,967	11.17%	34,009	10.84%
3	55,888	32.74%	45,258	31.65%	101,146	32.25%
4	63,232	37.04%	53,218	37.22%	116,450	37.12%
5	17,590	10.30%	12,814	8.96%	30,404	9.69%
6	2,558	1.50%	1,226	0.86%	3,784	1.21%
Total	170,696	100.00%	142,983	100.00%	313,679	100.00%



3.2.4.4 Grades 6-8

Table 3.2.4.4
Proficiency Level Distribution: Over 6-8 S601 Online

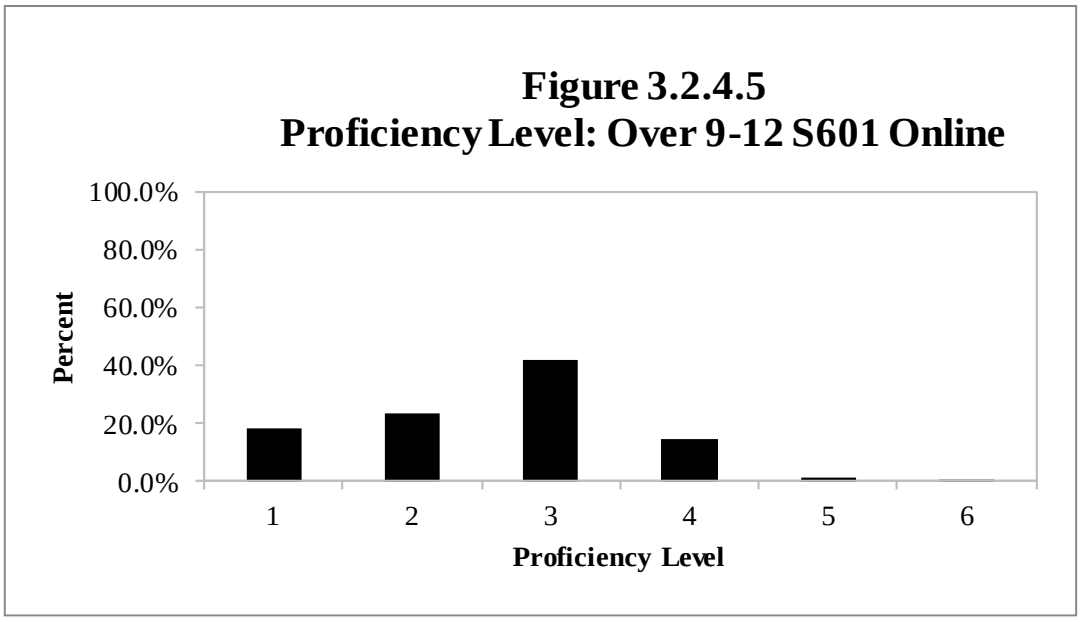
	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	14,423	12.21%	17,815	14.98%	20,301	17.82%	52,539	14.97%
2	26,499	22.43%	25,719	21.63%	24,464	21.48%	76,682	21.85%
3	58,653	49.64%	54,052	45.46%	48,171	42.29%	160,876	45.84%
4	17,797	15.06%	20,187	16.98%	19,765	17.35%	57,749	16.45%
5	745	0.63%	1,096	0.92%	1,175	1.03%	3,016	0.86%
6	43	0.04%	23	0.02%	35	0.03%	101	0.03%
Total	118,160	100.00%	118,892	100.00%	113,911	100.00%	350,963	100.00%



3.2.4.5 Grades 9-12

Table 3.2.4.5
Proficiency Level Distribution: Over 9-12 S601 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	23,730	18.01%	19,601	18.27%	13,754	17.84%	13,004	20.04%	70,089	18.39%
2	28,830	21.88%	24,966	23.27%	18,111	23.49%	17,570	27.08%	89,477	23.48%
3	55,941	42.46%	44,736	41.70%	32,516	42.17%	26,194	40.37%	159,387	41.83%
4	21,004	15.94%	16,248	15.14%	11,372	14.75%	7,397	11.40%	56,021	14.70%
5	2,149	1.63%	1,677	1.56%	1,298	1.68%	707	1.09%	5,831	1.53%
6	111	0.08%	56	0.05%	51	0.07%	9	0.01%	227	0.06%
Total	131,765	100.00%	107,284	100.00%	77,102	100.00%	64,881	100.00%	381,032	100.00%



4. Annual Updates of Validity Evidence

This section presents studies conducted as validity evidence for the WIDA ACCESS assessments. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), validity is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use. Particular interpretations for specified uses begin by specifying the construct the test is intended to measure. Rather than referring to distinct types of validity, the Standards refer to types of validity evidence. According to the Standards, the evidence can be based on (1) test content, (2) response processes, (3) internal structure, and (4) relation to other variables.

The validity evidence of the Standards is also observed in "A State's guidance to the U.S. Department of Education's Assessment Peer Review Process" document (Department of Education, 2018 https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf) to support states' use of ELP assessments for reviewing of validity evidence, as well as is linked to Assessment User Argument (AUA) to support the claims of validity of Online ACCESS assessment. WIDA structures its validity arguments using AUA model in lieu of the model highlighted in the *Standards for Educational and Psychological Testing.* AUA has similar topics; however, they are organized differently. Below is a short summary of each AUA claim. For the full AUA validity claims, please refer to WIDA Assessment User Argument document.

Claim 1 (Consequences): With the use of ACCESS, the intended decisions will have beneficial consequences for stakeholders, in terms of using ACCESS and the decisions made based on ACCESS.

Claim 2 (Decisions): Decisions based on ACCESS test results are made by individuals, in a timely manner, and affect a variety of stakeholders. Two types of decisions that are made based on ACCESS results are classification and programming decisions. The decisions take into consideration educational and societal values, and relevant laws, rules, and regulations, and they are equitable for the intended stakeholders.

Claim 3 (Interpretations): The interpretations of students' academic English language proficiency in four domains are *relevant* to the classification, placement and programming decisions; *sufficient*, in conjunction with additional information as outlined in state and local

policies, to make such decisions; *meaningful* with respect to the WIDA English Language Development (ELD**)** Standards; *generalizable* to the academic English language used in K–12 instructional settings, and *impartial* to all students.

Claim 4 (Assessment records: Scores): ACCESS scores are consistent across different aspects of test administration, different test tasks, and different groups of students. Test forms and metrics accurately represent the construct being measured and result in expected test taker performances.

4.1 Standards

4.1.1 Test Content

The relationship between the content of a test and the construct to measure is called content validity. Test content includes the themes, wording, and format of the items, tasks, or questions on a test. Administration and scoring may also be part of the content. Empirical or logical evidence can show how appropriate the content reflects the domain as we interpret test scores.

4.1.2 Response Processes

Empirical analysis of how test takers process tests provide evidence of the nature between performance and the construct. Examples of this validity include analyzing individual item responses, different response processes in answering questions by subgroups or evaluating test-takers' performance.

4.1.3 Internal Structure

Validity related to internal structure indicates how test items/components agree with the construct we base for the score interpretation. The internal structure of the construct can be unidimensional or contain multidimensional components.

4.1.4 Relation to Other Variables

The interpretation of the test scores with an external indicator provides valuable validity evidence. We often ask how accurately the test score predicts the criterion variable. The test criterion validity has two different validities: concurrent and predictive validity. Predictive validity is how accurately test scores predict the future performance of criterion scores. Concurrent validity indicates how test scores relate to criterion scores at the same time.

4.2 Annual Validity Studies

4.2.1 Detection of Multiple Group Differential Item Functioning for Students with Disabilities Taking an English Language Proficiency Assessment

This study aims to explore differential item functioning (DIF) by disability groups in a large-scale alternate English assessment for students with cognitive and physical disabilities. The preliminary step in most DIF methods is to designate a reference group and focal group(s) and then detect item bias towards one group. However, it is challenging to conduct DIF studies in the context of this alternate assessment because there is no reference group. Furthermore, it is even more challenging to conduct DIF studies when the sample size is small and unequal, the ability distributions differ across groups, the test length is short, and items follow polytomous scoring. In this circumstance, simultaneous DIF across all groups (Kim et al., 1995) might be preferred over multiple pairwise DIFs (Ellis and Kimmel, 1992) to address no reference group and multiple groups challenge while controlling Type I error rate. This presentation illustrates how a comparison of item parameters (Thissen, Steinberg, & Wainer, 1993) and item response functions (IRF) (Wainer, 1993) of multiple groups work in real-world data. Furthermore, the initial insights into items exhibiting DIF and considerations for developing inclusive items for students with disabilities will be shared.

The study addressed the following research questions:

- How do various DIF methods perform with polytomous items among multiple groups of disability across five FT test forms?
- Is there any difference between Item Response Theory and Classical Test Theory methods in terms of flagging?

The results of this study support the following conclusions:

- Characteristics of individual forms are different in terms of disability group, sample size, and item difficulty. Even though many items were flagged as DIF, they were not consistently flagged across forms.

- Some field test items are somewhat consistently flagged as DIF needs to be reviewed by content experts.
- Field test items do not consistently favor any of the considered disability groups.
- DIF results are highly dependent on the DIF method and sample.
- Reading and Writing are more prone to DIF because they are more labor-intensive due to disability as compared to Listening and Speaking
- LR and MH are more sensitive due to the small sample size compared to Rasch, RMSD, and Lasso.

A report on this study has been made available on the Secure Portal: https://wida.wisc.edu/resources/alt-access-field-test-differential-item-functioning-analysis-report

4.2.2 English Learners' Use of Universal Tools: Interview Study

English learners (ELs) comprise approximately 10% of kindergarten to grade 12 students in U.S. public schools, with about 15% of ELs identified as having disabilities. English language proficiency (ELP) assessments must adhere to universal design principles and incorporate universal tools, designed to increase accessibility for all ELs, including those with disabilities. The purpose of the two-phase mixed- methods study was to examine the extent Grades 1–12 ELs with and without disabilities activated universal tools during an online ELP assessment: Color Overlay, Color Contrast, Help Tools, Line Guide, Highlighter, Magnifier, and Sticky Notes. In Phase 1, analyses were conducted on 1.25 million students' test and telemetry data (record of keystrokes and clicks). Phase 1 findings showed that ELs activated the Line Guide, Highlighter, and Magnifier more frequently than others. The tool activation rate was higher in listening and reading domains than in speaking and writing. A significantly higher percentage of ELs with disabilities activated the tools than ELs without disabilities, but effect sizes were small.

Phase 2 aimed to examine students' rationale for activating the universal tools via interviews. It addressed the below questions:

1) To what extent do ELs, with and without disabilities, activate the universal tools in an ELP assessment?
2) What were ELs' rationales for activating the universal tools?

A total of 55 Grades 4-12 ELs were interviewed for about 20 minutes after test administration (Grades 1- 3 were not included as they complete the writing test via paper). Fifteen of the participants had disabilities. All had a primary disability of *specific learning disability* and four students had a secondary disability of *speech/language impairment*. Interviews were video-recorded and transcribed. Students' responses were analyzed according to (1) the tools they reported having used during the test, (2) their rationale for tool use, and (3) suggestions for improving the tools.

Findings revealed no meaningful differences in tool use between ELs with and without disabilities as both groups activated the same tools most often (highlighter, magnifier, and line guide). When students activated the universal tools, they used them in expected ways. For example, the highlighter and line guide allowed test takers to focus their attention on specific words on the screen. Students had insightful suggestions for further enhancing the quality of the universal tools, such as having multiple colors for the Highlighter, allowing students to choose zoom ratio in Magnifier, or an embedded dictionary tool.

Findings have been published in a special issue of *Language Testing*: https://journals.sagepub.com/doi/abs/10.1177/02655322221149009

4.2.3 Impact of Ability Range Restriction on Item Characteristics in Multistage Adaptive Testing

Student ability range and item difficulty range can influence the range of the possible fit statistics, although not many studies on this topic exist in the literature. Infit and outfit statistics are impacted by all observations. Outfit is outlier sensitive, that is, is sensitive to unexpected observations by students on items that are relatively easy or very hard for them (Linacre, 2002). If a certain ability range is missing in the item responses, it might affect the calculation of outliers. Infit statistics are average standardized residuals weighted by information (variance) (Linacre, 2002). Infit statistics are more sensitive to unexpected patterns of observations by students on items that are roughly targeted to them. The total variance of item performance might be reduced or increased depending on the test takers' ability range, which therefore may impact infit results.

This study questions how different conditions of ability range in the FT item calibration in MST affect item parameters and infit and outfit statistics in the Rasch model. To investigate this question, we use empirical and simulated data in various conditions of 1) ability ranges of examinees, 2) sampling design, and 3) sample sizes.

This study explores how differences in test takers' ability range affect item characteristics such as item difficulty, point-measure correlation, and fit statistics in multistage adaptive testing (MST) under the Rasch model. In MST, students are presented with sets of different items based on their ability levels. For test refreshment and item security purposes, new items need to be added to the operational test periodically. Depending on how we administer new field test (FT) items (whether each new field test item is administered to targeted ability ranges of test takers or is given to broader ability ranges), item placement might affect item characteristics in the calibration.

Routing methods (current ACCESS routing vs MFI routing) showed different performance across clusters and domains in terms of SE, RMSE, and item parameter recovery. Correlation between initial abilities and the estimated abilities varied depending on clusters and domains. Regarding module exposure, MFI sometimes did not choose certain tier folders because there was not much difference of information between tier folders. The main takeaway from this study is that random FT folder assignment is shown best in recovering the true item difficulty values. It means our current FT assignment is better than targeted tier FT assignment but still somewhat restricted range of ability. The full random assignment of FT items is closer to the true difficulty values.

This study was presented at the 2023 meeting of the National Council of Measurement in Education. A report on this study has been made available on the Secure Portal: https://wida.wisc.edu/resources/impact-ability-range-restriction-item-characteristics-access-multistage-adaptive-testing

4.2.4 WIDA Standards-ACCESS Alignment

In response to concerns expressed by federal Department of Education peer reviewers, we conducted an alignment study between the WIDA ELD Standards and ACCESS. A panel of

experts in ELD standards was convened to examine the degree of alignment following procedures developed by Webb (1997)[7], which defines alignment in terms of match, depth, and breadth. We investigated alignment in two dimensions: domain by proficiency level, and standard across domains. The domain-by-proficiency-level alignment analyses showed that the match and depth criteria were met for almost all proficiency levels across the domains. However, the breadth criteria had mixed results: Range was generally met for Listening and Reading, though not for Speaking and Writing; and Balance was generally not met. For the alignment analysis by standard across domains, match and depth criteria were generally strongly or moderately met. Limited range, which is one of the breadth criteria, was found for most clusters in Language of Mathematics and Language of Social Studies. Balance, which indicates any emphasis on assessed targets, was generally limited or moderate, and may reflect the test design—some standards are intended to be emphasized more than others.

A full report on the study was submitted to the U.S. Department of Education for peer review.

4.2.5 WIDA Standards Correspondence

In August 2023, WIDA released a technical paper, "WIDA Correspondence Mapping of the Match, Breadth, Consistency, and Depth of Language Opportunities in State K–12 English Language Arts, Mathematics, Science, and Social Studies Standards."

The goal of the study was to ensure that the WIDA Key Language Uses and Language Expectations would be flexible enough to fit with many different content areas and types of standards, whether "multistate" or individual in nature. The technical paper addresses four research questions (RQs):

- RQ1: What is the degree of match between state academic content standards and the WIDA Key Language Uses?

- RQ2: What is the breadth of coverage by Key Language Uses in state academic content standards?
- RQ3: What is the balance of representation of Key Language Uses in state academic content standards?

- RQ4: What is the depth of linguistic complexity in the match between the WIDA Language Expectations and WIDA Proficiency Level Descriptors?

The analyses reported in this paper were carried out during 2019-2020 while developing the WIDA ELD Standards Framework, 2020 Edition. The findings were later cross-checked and updated using the Fall 2022 versions of academic content standards from all WIDA consortium member states.

Findings establish a direct relationship between the four WIDA Key Language Uses (and their instantiation in grade-level cluster Language Expectations) and language uses found in state academic content standards and illustrate the relationship between the grade-level cluster Language Expectations and Proficiency Level 5 of the Proficiency Level Descriptors.

Findings underscore the adaptability of the WIDA ELD Standards Framework, its comprehensive coverage, and its alignment with academic content standards. Identifying correspondences ensures alignment between ELP standards and academic content standards, ensuring compliance, effective instruction for English language learners, and fostering student achievement. This technical paper provides evidence to support WIDA consortium member SEAs in complying with the Every Student Succeeds Act of 2015 and related U.S. Department of Education peer review guidance (Peer Review Critical Element 1.2).

WIDA presented the study findings previously at the 2022 WIDA Board Meeting. The technical paper is shared as WCER Working Paper No 2023-3 and was also distributed to all WIDA Consortium member SEAs. It is available at https://www.wcer.wisc.edu/publications/abstract/wcer-working-paper-no-2023- 3.

5. Reliability

In accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014), when interpreting test scores, it is important to evaluate their reliability, as the interpretation of test scores depends on the assumption that students exhibit some degree of consistency in their scores across independent administrations of the same testing procedure. We expect that students mastering the domain will consistently perform well, and those who have not mastered the domain will consistently perform less well, regardless of the sample of items and tasks used to assess students. Furthermore, because we assume that all items and tasks on such a test measure some aspect of the domain of interest, we expect that students will perform consistently across different items and tasks measuring the same ability within the test. Therefore, it is important to evaluate the degree to which students' test scores are consistent across replications of the same testing condition.

However, different samples of performances from the same student are rarely identical. A student's responses to sets of test items or tasks vary from one sample of test items or tasks targeting the domain to another, and from one occasion to another, even under strictly controlled conditions. In addition, different raters may award different scores to the same student performance on a test task. These sources of variation are reflected in the students' scores. Therefore, it is important to evaluate the extent to which differences in students' test scores reflect true differences in the knowledge, skills, or ability being tested, rather than fluctuations due to chance.

The reliability of the test scores depends on how much the scores vary across replications of the testing procedure, and analyses of reliability depend on the types of variability likely to be of concern in the testing procedure. There are several ways to collect reliability data and to estimate reliability, some of which depend on the exact nature of the measurement, the intended use of the test scores, the assessment design, and the potential sources of measurement error that might contribute to inconsistency in students' scores across different test administrations.

The reliability information presented in this section is organized to be in compliance with Critical Element 4.1 of the Every Student Succeeds Act Peer Review requirements (U.S. Department of Education, 2018) and follows the guidelines of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). We present

information regarding the reliability of the domain scale scores first, followed by information about the reliability of the composite scale scores.

Policy makers in states and districts use ACCESS Listening, Reading, Writing, and Speaking tests to determine the English language proficiency of students based on their scores in each of the four domains. Therefore, the main concern in interpreting these scores is how consistent the scores would be over replications of the same testing procedure. We use **internal consistency reliability statistics** to address this question (Section 5.1).

Additionally, for the Writing and Speaking domains, because having different raters evaluate the same students' responses to tasks may result in inconsistent scoring, a potential source of variation of those scores is the rater. In Section 5.2, we report the **interrater agreement** rates that the raters achieved when evaluating students' responses to the Writing and Speaking tasks. We can use these statistics to determine how consistent the students' scores would have been if different raters had evaluated their responses. Since we use an item response theory (IRT)–based method to estimate students' **latent scores** (i.e., test scores based on variables that we cannot see or directly measure but which we can infer mathematically through advanced statistical techniques by using students' scores on variables that we can observe), we also examine the amount of **measurement error** in students' scores using the **conditional standard error of measurement** (CSEM) (Section 5.3). Lastly, in Section 5.4, we evaluate the reliability of the classifications of students into WIDA proficiency levels based on their domain scores (the most important interpretation of the test scores) in terms of the **accuracy and consistency** of the classification decisions made. In each subsection, we present detailed descriptions of the methods, data sources, and procedures.

Policy makers in states and districts use ACCESS **composite scale scores** to describe the English language proficiency of students in the respective composites. Therefore, the most important concern in interpreting these scores is how consistent the scores would be over replications of the same testing procedure. We use internal consistency reliability statistics to address this question, and in Section 5.5 we provide the results. In addition, in Section 5.6, we examine the CSEM of these scores. Lastly, in Section 5.7, we evaluate the reliability of the classifications in terms of the accuracy and consistency of the decisions made about students' levels of English language

proficiency based on their composite scale scores. In each subsection, we present detailed descriptions of the methods, data sources, and procedures.

Internal Consistency Reliability Statistics

One way to evaluate the consistency of students' test scores across test administrations is to examine how the students would have performed on alternate forms of the same test (i.e., **parallel test form reliability**). Given our assumption that the ability the test measures is constant for each student over two administrations of alternate forms, the more variation found across the two administrations, the more evidence for lower reliability. The **measurement error** represents the sources of inconsistency across the two administrations, taken together. We consider measurement error to be random and to occur by chance. For example, there may be some construct-irrelevant knowledge and/or skills that some items or tasks measure that affect students' scores but are not part of the ability that the test intends to measure.

Unless students take two alternate versions of the same test, we cannot calculate test score reliability directly. Thus, we usually estimate it from student responses to a single form of the test. Methods employed to estimate reliability using test scores from a single test administration are based on classical test theory and are referred to as estimates of **internal consistency**. An internal consistency reliability statistic is a useful estimate of alternate-forms reliability, providing an estimate of the consistency of students' performances across items and tasks within a test. The most common index of internal consistency reliability is **Cronbach's coefficient alpha** (Cronbach, 1951), which is a lower-bound estimate of test reliability. Conceptually, we think of Cronbach's coefficient alpha as the correlation obtained between performances on two halves of the same test if every possible way of dividing the test items and tasks in two were attempted. Because Cronbach's coefficient alpha is a correlation of students' performances on all possible pairs of test items and tasks, it may be low if some items or tasks are measuring something other than what most of the other items and tasks are measuring (and thus leading to inconsistent student performances). In this way, Cronbach's coefficient alpha expresses how well the items and tasks on a test appear to measure the same ability. The Cronbach's coefficient alpha of internal consistency ranges from 0 to 1. If students achieve their scores by a completely random process (i.e., their scores are not correlated or share no covariance), then the reliability

estimate is very close to 0. On the other hand, if students' scores are perfectly consistent (i.e., their scores have high covariances), then the internal consistency coefficient will approach 1.

While there is no one set of criteria that the testing community uses when interpreting Cronbach's coefficient alpha values, from time to time, researchers have proposed various arbitrary criteria that one could apply. Initially, Cronbach (1951) argued that it was 'desirable' to have a high alpha value for an instrument that test developers were using to report individual scores since the scores on that instrument needed to be interpretable, and that would require a high alpha value. Later, Nunnally (1978) suggested that researchers should consider a value of 0.70 as an acceptable lower limit if they were engaged in the early stages of research (e.g., when developing a scale). Today, it has become common practice to cite Nunnally's suggested 0.70 criterion as a minimum acceptable lower limit for this value for all types of research. However, in so doing, researchers ignore Nunnally's more nuanced guidance: If researchers were engaged in basic research, Nunnally advised that they should use a higher cut-off value (i.e., 0.80 or higher), and those engaged in applied research should use a much higher cut-off value (0.90 or higher) (Lance et al., 2006). Since Nunnally's time, some researchers have suggested even more nuanced interpretations of various alpha values. For example, George and Mallery (2003) proposed the following interpretations: "≥ 0.90 – Excellent, ≥ 0.80 – Good, ≥ 0.70 – Acceptable, ≥ 0.60 – Questionable, ≥ 0.50 – Poor, and ≤ 0.50 – Unacceptable" (p. 231). Clearly, there is little consensus among the experts in their views of what the acceptable lower limit of the Cronbach's coefficient alpha value should be, or for that matter, how one should interpret various values. This lack of consensus led the authors of the *Standards for Educational and Psychological Measurement* (2014) to conclude, "The choice of [reliability/precision] estimation and the minimum acceptable level for any index remain a matter of professional judgment" (p. 41). For the purposes of this report then, WIDA has made the decision that within the domains of Listening, Reading, and Speaking, an alpha value of ≥ 0.80 is acceptable, while an alpha value of ≥ 0.65 is acceptable for the Writing domain.

Reliability statistics such as the Cronbach's coefficient alpha of internal consistency are affected by two factors: (1) the number of test items or tasks, and (2) the total number of score points students achieve. That is, all things being equal, the greater the number of items or tasks measuring the same ability there are on the test, the higher the internal consistency reliability statistics. Additionally, because reliability statistics refer to the consistency of scores *for a group*

of students, the distribution of that specific group's ability measures affects these statistics. If the students in the group are nearly equal in the ability that the test measures (i.e., their scores are concentrated in the center of the ability distribution), small changes in their scores can easily change their relative positions in the group. Consequently, the internal consistency reliability statistics will be low. In this case, the statistic may be telling us more about the group of students tested than about the test itself. On the other hand, if the students in the group differ widely in the ability that the test measures (i.e., their scores are distributed across the ability continuum), small changes in their scores will not affect their relative positions in the group as much, and the internal consistency reliability statistics will be higher. Therefore, reliability can be as much a function of the performance of test items and tasks as of the performance of the sample of students tested. That is, the exact same test can produce widely disparate reliability indices based on the ability distribution of the group of students. This means, in turn, that when interpreting estimates of internal consistency, it is wise to keep in mind the specific set of test items and tasks and the distribution of ability measures in the group of students used in the estimation.

Interrater Agreement

The behavior of raters is a potential source of variance in students' scores for the productive domains of ACCESS (i.e., Writing and Speaking). ACCESS scoring procedures and rater training and quality control monitoring processes are described elsewhere in this report (see Part 1, Section 3.2.2). In Section 5.2, we report the **interrater agreement rates** for the scoring of students' responses to the Writing and Speaking tasks. These values reflect how consistent the students' scores would be if different groups of raters scored their responses. Additionally, in this section of the report we present a detailed description of the methods, data sources, and procedures we used when calculating interrater agreement rates.

Measurement Error

In addition to evaluating test score reliability in terms of estimates of internal consistency, we can calculate the amount of measurement error in students' test scores in two different ways. One way is to hypothesize that there is an error-free measure of each student's true ability, referred to as the **true score** in classical test theory. The true score is a theoretical value, so it is not a known quantity. Rather, we view it as the hypothetical average score over repeated replications of the same testing condition (Livingston, 2018, p. 9). Under the assumptions of classical test theory,

the **error of measurement** over a replication of a testing condition provides an estimate of the amount of variability from students' true scores that we would expect. In practical testing contexts, it is generally not possible to replicate a testing condition (i.e., have students take the same test form multiple times), so it is not possible to estimate the standard error of each student's score using a repeated measures design. Instead, we calculate the average error of measurement over the population of students who take the test, and then we use that as an indication of the amount of variation in any individual student's score that we would expect. Classical test theory refers to this average as the **standard error of measurement** (SEM), which provides an indication of how much students' scores differ from their true scores, on average, on the raw score metric. Because it is a standard deviation of the distribution of errors of measurement, we can construct a **confidence interval** to indicate how the errors of measurement are affecting the scores. Test scores with large SEMs pose a challenge to the interpretation of the reliability of any single test score.

A second way to address the impact of measurement errors on students' test scores is to estimate the SEM for specific scores using IRT. IRT addresses reliability using the **test information function**, which indicates the precision with which we can use student performances on items and tasks to estimate the **latent** (i.e., true) **ability** of each student (i.e., **latent scores**). The square root of the inverse of the information function at any point on the latent ability distribution is the **conditional standard error of measurement** (**CSEM)**. The CSEM provides information about the amount of error we would expect in any student's score at that point on the underlying latent ability scale, which IRT refers to in terms of the **latent score metric** (i.e., the IRT metric for expressing student ability, as opposed to the raw score metric). In addition, by using IRT, we can estimate indices analogous to traditional reliability coefficients such as Cronbach's coefficient alpha from the test information function and the distribution of the latent scores in the same student population.

Classification Accuracy and Consistency

One of the main purposes of the WIDA ACCESS program is to identify the English language proficiency levels of students with respect to the WIDA ELD Standards. Because of the emphasis on the classification of student performance into six WIDA proficiency levels, it is important to know how consistently ACCESS scores do indeed classify students into those

proficiency levels (American Educational Research Association et al., 2014). The questions that we want to answer are different from the questions that the reliability coefficient answers. Instead of looking at the reliability of a specific student score, we want to know the consistency of the decisions we make when we use students' test scores to classify them into a smaller number of proficiency levels. One way to approach this question is to estimate the degree to which the classification decisions we are making based on the students' **observed test scores** agree with the classification decisions we would make based on students' **theoretical true scores**. This estimate is known as **decision accuracy**. A second way to approach this question is to estimate the degree to which the classification decisions we are making based on the students' test scores agree with the classification decisions we would make based on students' scores on an alternate form of the test. This estimate is known as **decision consistency**.

5.1 Reliabilities of the Domain Scores

Listening and Reading

Internal consistency statistics based on classical test theory are applicable only for a fixed-length test where all students take the same set of test items (Thissen, 2000). For the Listening and Reading tests, which are computer adaptive, we cannot compute traditional internal consistency reliabilities because not all students take the same set of items. We estimate the reliabilities of students' domain scale scores for Listening and Reading by grade-level cluster using an IRT-based **marginal reliability method** that Thissen (2000) derived. Unlike the traditional internal consistency statistics that are based on students' raw scores, the marginal reliability method for calculating reliability uses students' domain scale scores and the distribution of the students' domain scale scores on the theta scale (i.e., **domain theta scores**) in its estimation. However, we can interpret the marginal reliability coefficient like other traditional internal consistency coefficients such as Cronbach's coefficient alpha (Thissen, 2000).

The formula for calculating an IRT-based marginal reliability coefficient using the method that Thissen (2000) developed is

$$\bar{\rho} = \frac{\sigma_\theta^2 - average(CSEM_{observed}^2)}{\sigma_\theta^2}$$

where

$\bar{\rho}$ is the average reliability

σ_θ^2 is the variance of the distribution of the students' domain theta scores

$CSEM_{observed}^2$ is the squared observed CSEM for each student's domain theta score.

We can calculate the IRT-based marginal reliability coefficient directly (Thissen, 2000); however, it is computationally intensive. Since this estimate is equivalent to the **Rasch student separation reliability coefficient** (Linacre, 1999), which is regularly reported as part of the output from a Winsteps analysis, for purposes of efficiency WIDA chose to report the Rasch student separation reliability coefficients as the test score reliability estimates for the Listening and Reading domains. The Rasch student separation reliability coefficient is an estimate of the ratio of "true measure variance" to "observed measure variance" (Linacre, 1999). The student

separation reliability coefficient answers these questions: How consistent are the students' relative positions in the group tested, as indicated by their domain scale scores? How reproducible is the student ability measure order of this sample of students for this set of items? The more the students differ in ability, the less likely that small changes in their domain scale scores will affect their relative positions in the group, and the higher the student separation reliability coefficient will be. Thus, to obtain high student separation reliability, a wide sample of student ability in the domain (i.e., a large student ability range) and/or low measurement error (i.e., a test containing many items) is required (Linacre, 2020). Student separation reliabilities can range from 0.00 to 1.00. A student separation reliability < .80 implies that the test may not be sensitive enough to distinguish between high- and low-performing students, and thus more items may be needed (Linacre, 2020). To obtain these values, we used the item parameters and population student data as inputs for the Winsteps program.

In the following tables, which present test score reliability information for the Listening and Reading domains, we provide the Rasch student separation reliability coefficients that are based on students' ACCESS Online domain theta scores. For these two domains, the first table reports the Rasch student separation reliability coefficient (labeled as 'Rasch Student Separation Reliability Coefficient' in the table) for all students in each grade-level cluster. Each row in the table represents a grade-level cluster, and values for the numbers of students, numbers of items, and the student separation reliability estimate are provided based on students' domain theta scores in each grade-level cluster. The second table for each domain provides the same information for the population of female students and for the population of male students. The third table provides information by ethnicity, for Hispanic and for non-Hispanic students, and the fourth table provides information for the population of students who have an individualized education plan (IEP).

For Listening, the Rasch student separation reliability coefficients based on the domain theta scores for all students ranged from 0.85 to 0.88 across the grade-level clusters. The Rasch student separation reliability coefficients ranged from 0.85 to 0.89 for male students; 0.85 to 0.87 for female students; 0.85 to 0.88 for Hispanic students; 0.83 to 0.88 for non-Hispanic students; and 0.79 to 0.88 for students with an IEP.

For Reading, the Rasch student separation reliability coefficients based on the domain theta scores for all students ranged from 0.84 to 0.90 across the grade-level clusters. The Rasch student separation reliability coefficients ranged from 0.84 to 0.90 for male students; 0.84 to 0.90 for female students; 0.80 to 0.90 for Hispanic students; 0.87 to 0.91 for non-Hispanic students; and 0.79 to 0.87 for students with an IEP.

Writing and Speaking

Cronbach's coefficient alpha is widely used as an estimate of reliability, particularly for the internal consistency of test items and/or tasks, and this statistic is appropriate for calculating the reliabilities of students' scores from the administration of the fixed forms of the Writing and Speaking tests. Conceptually, we can think of it as the correlation obtained between students' performances on two halves of the Writing or Speaking test if every possible way of dividing the test tasks in two were attempted. Thus, Cronbach's coefficient alpha may be low if some tasks are measuring something other than what the majority of the tasks are measuring. In this way, Cronbach's coefficient alpha expresses how well the tasks on a test appear to measure the same ability.

The formula for calculating Cronbach's coefficient alpha for the fixed forms of the Writing and Speaking tests is

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

where

n = the number of tasks

σ_i^2 = the variance of students' raw scores on task i

σ_t^2 = the variance of students' total raw scores.

For the Writing and Speaking tests, tables in this section also present the SEM, a single value for estimating the errors of measurement in students' raw scores calculated using a classical test theory-based approach. It is a function of two statistics: (1) the Cronbach's coefficient alpha calculated using students' raw scores on the test, and (2) the (observed) standard deviation (SD)

of the students' total raw scores. It is on the raw score metric. The Cronbach's coefficient alpha is calculated as

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

Since the SEM is an estimate of the standard deviation of the distribution of measurement errors, we can use the SEM to create a band around a student's observed raw score. Under the assumption that the error of measurement follows a normal distribution, the student's true score would lie with a certain degree of probability within this band. Statistically speaking, then, there is an expectation that a student's true raw score has a 68% probability of falling within the band extending from the observed score minus 2 SEMs to the observed score plus 2 SEMs. Since SEMs are expressed on the raw score metric, it is wise to keep the range of the possible raw score distribution in mind when interpreting the SEM. For example, if the Online Writing test has a possible raw score range of 0 to 18 and one SEM equals 2 score points, and if a student receives a score of 10 on the test, we know with 95% certainty that the student's true score lies somewhere between a raw score of 8 and 12 (i.e., 10 minus, or plus, 2 SEMs). Similarly, if one SEM equals 1 score point, we would say with 68% certainty that the student's true score lies between 9 and 11 (i.e., 10 minus, or plus, 1 SEM). The smaller the value of the SEM, the more precise the test scores will be.

The range of total possible raw score points for the Writing forms is 0 to 18. The ranges of total possible raw score points for the Speaking forms are 0 to 6 for Tier Pre-A, 0 to 18 for Tier A, and 0 to 24 for Tier B/C. As described in Section 2.3, given the semi-adaptive nature of the Speaking test, students taking the Tier B/C form of the Speaking test receive an additional 2 points for each of the three PL 1 tasks that they did not take. Consequently, the officially reported total raw score points for the Speaking Tier B/C form range from 6 to 30. However, since we computed the Cronbach's coefficient alpha for the Speaking Tier B/C form using students' raw scores on the six Speaking tasks that the students actually took, the total possible raw score points reported in the SEM tables in this section range from 0 to 24—that is, without the six free points added to the total possible raw score points.

The tables in the next section that present reliability information for the Writing and Speaking tests report the number of tasks, the Cronbach's coefficient alphas, and the SEMs for all students and for subgroups as the Every Student Succeeds Act Peer Review requires, thus facilitating the

comparison of the reliability estimates computed based on the performance of individual subgroups to those computed based on the performance of all students. For these domains, the first table provides the Cronbach's coefficient alphas and the SEMs for all students based on their raw scores. Each row in the table represents a specific grade-level cluster and test form. For each form, the tables provide the numbers of students, numbers of tasks, total possible raw score points, Cronbach's coefficient alpha, and SEM. The second table for each domain provides the same information for the population of female students and for the population of male students. The third table provides information by ethnicity, for Hispanic and for non-Hispanic students, and the fourth table provides information for the population of students who have an IEP.

Note that students' Writing reported scores are based on their performances on only two tasks starting with Online Series 501. Therefore, the Cronbach's coefficient alpha for the Writing domain may be lower than when estimated based on student performances on three tasks, as in earlier series.

Writing Tier A: The Writing Tier A Cronbach's coefficient alphas computed based on the raw scores for all students ranged from 0.86 to 0.89. The Writing Tier A Cronbach's coefficient alphas ranged from 0.86 to 0.89 for male students; 0.86 to 0.89 for female students; 0.85 to 0.89 for Hispanic students; 0.85 to 0.88 for non-Hispanic students; and 0.82 to 0.87 for students with an IEP.

Writing Tier B/C: The Writing Tier B/C Cronbach's coefficient alphas computed based on the raw scores for all students ranged from 0.65 to 0.78. The Writing Tier B/C Cronbach's coefficient alphas ranged from 0.66 to 0.78 for male students; 0.64 to 0.76 for female students; 0.66 to 0.79 for Hispanic students; 0.65 to 0.73 for non-Hispanic students; and 0.64 to 0.83 for students with an IEP.

Speaking Tier Pre-A: The Speaking Tier Pre-A Cronbach's coefficient alphas computed based on the raw scores for all students ranged from 0.86 to 0.87. The Cronbach's coefficient alphas ranged from 0.86 to 0.87 for male students; 0.84 to 0.87 for female students; 0.85 to 0.87 for Hispanic students; 0.84 to 0.88 for non-Hispanic students; and 0.84 to 0.91 for students with an IEP.

Speaking Tier A: The Speaking Tier A Cronbach's coefficient alphas computed based on the raw scores for all students ranged from 0.84 to 0.85. The Cronbach's coefficient alphas ranged

from 0.84 to 0.85 for male students; 0.83 to 0.85 for female students; 0.84 to 0.85 for Hispanic students; 0.80 to 0.82 for non-Hispanic students; and 0.79 to 0.85 for students with an IEP.

Speaking Tier B/C: The Speaking Tier B/C Cronbach's coefficient alphas computed based on the raw scores for all students ranged from 0.84 to 0.88. The Cronbach's coefficient alphas ranged from 0.84 to 0.88 for male students; 0.83 to 0.87 for female students; 0.84 to 0.88 for Hispanic students; 0.82 to 0.85 for non-Hispanic students; and 0.83 to 0.88 for students with an IEP.

5.1.1 Listening

Table 5.1.1.1

Reliabilities of Domain Scores: List S601 Online

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	213,228	54	0.88
2-3	415,059	54	0.88
4-5	364,830	54	0.85
6-8	413,790	54	0.85
9-12	445,815	54	0.86

Table 5.1.1.2

Reliabilities of Domain Scores: List S601 Online by Gender

		Female		Male	
Cluster	No. of Items	No. of Students	Rasch Student Separation Reliability Coefficient	No. of Students	Rasch Student Separation Reliability Coefficient
1	54	88,270	0.87	95,972	0.88
2-3	54	170,376	0.87	187,537	0.89
4-5	54	143,943	0.85	168,153	0.86
6-8	54	159,018	0.85	194,398	0.85
9-12	54	168,874	0.85	212,351	0.86

Table 5.1.1.3

Reliabilities of Domain Scores: List S601 Online by Ethnicity

		Hispanic		Other	
Cluster	No. of Items	No. of Students	Rasch Student Separation Reliability Coefficient	No. of Students	Rasch Student Separation Reliability Coefficient
1	54	141,540	0.88	65,407	0.87
2-3	54	274,923	0.88	127,928	0.88
4-5	54	246,351	0.85	102,377	0.83
6-8	54	290,889	0.85	101,054	0.84
9-12	54	310,683	0.85	110,380	0.83

Table 5.1.1.4

Reliabilities of Domain Scores: List S601 Online by IEP Status

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	18,922	54	0.88
2-3	41,409	54	0.88
4-5	44,160	54	0.84
6-8	59,855	54	0.81
9-12	58,205	54	0.79

5.1.2 Reading

Table 5.1.2.1

Reliabilities of Domain Scores: Read S601 Online

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	218,896	72	0.84
2-3	423,327	72	0.87
4-5	363,103	72	0.89
6-8	416,954	72	0.89
9-12	440,071	72	0.90

Table 5.1.2.2

Reliabilities of Domain Scores: Read S601 Online by Gender

		Female		Male	
Cluster	No. of Items	No. of Students	Rasch Student Separation Reliability Coefficient	No. of Students	Rasch Student Separation Reliability Coefficient
1	72	89,959	0.84	99,000	0.84
2-3	72	172,454	0.87	192,039	0.87
4-5	72	142,330	0.88	168,283	0.89
6-8	72	158,872	0.89	196,835	0.89
9-12	72	165,780	0.90	210,611	0.90

Table 5.1.2.3

Reliabilities of Domain Scores: Read S601 Online by Ethnicity

Cluster	No. of Items	Hispanic		Other	
		No. of Students	Rasch Student Separation Reliability Coefficient	No. of Students	Rasch Student Separation Reliability Coefficient
1	72	145,517	0.80	66,870	0.87
2-3	72	280,552	0.86	130,212	0.88
4-5	72	245,349	0.88	101,616	0.89
6-8	72	293,302	0.88	101,406	0.89
9-12	72	307,493	0.90	107,841	0.91

Table 5.1.2.4

Reliabilities of Domain Scores: Read S601 Online by IEP Status

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	19,703	72	0.79
2-3	42,667	72	0.83
4-5	44,354	72	0.87
6-8	61,160	72	0.85
9-12	57,886	72	0.87

5.1.3 Writing

Table 5.1.3.1

Reliabilities of Domain Scores: Writ S601 Online

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	A	205,653	2	0-18	0.86	1.15
	B/C	22,197	2	0-18	0.72	1.05
2-3	A	137,223	2	0-18	0.88	1.14
	B/C	308,682	2	0-18	0.78	1.19
4-5	A	86,447	2	0-18	0.89	1.04
	B/C	286,266	2	0-18	0.70	1.23
6-8	A	173,643	2	0-18	0.88	1.08
	B/C	254,635	2	0-18	0.70	1.03
9-12	A	172,511	2	0-18	0.89	1.04
	B/C	283,735	2	0-18	0.65	1.25

Table 5.1.3.2

Reliabilities of Domain Scores: Writ S601 Online by Gender

				Female			Male		
Cluster	Tier	No. of Tasks	Total Possible Raw Score Points	No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	A	2	0-18	83,990	0.86	1.15	92,874	0.86	1.13
	B/C	2	0-18	9,511	0.72	1.03	9,795	0.72	1.07
2-3	A	2	0-18	52,883	0.88	1.13	65,024	0.88	1.14
	B/C	2	0-18	128,777	0.76	1.16	136,195	0.78	1.19
4-5	A	2	0-18	31,921	0.89	1.05	42,162	0.89	1.03
	B/C	2	0-18	114,225	0.67	1.22	129,728	0.71	1.24
6-8	A	2	0-18	63,484	0.87	1.09	84,026	0.88	1.08
	B/C	2	0-18	99,491	0.67	1.02	117,743	0.72	1.04
9-12	A	2	0-18	60,833	0.88	1.06	86,952	0.89	1.03
	B/C	2	0-18	110,819	0.64	1.26	130,999	0.66	1.26

Table 5.1.3.3

Reliabilities of Domain Scores: Writ S601 Online by Ethnicity

Cluster	Tier	No. of Tasks	Total Possible Raw Score Points	Hispanic			Other		
				No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	A	2	0-18	141,829	0.85	1.15	57,578	0.85	1.14
	B/C	2	0-18	9,160	0.73	1.07	12,405	0.69	1.04
2-3	A	2	0-18	99,951	0.88	1.14	32,170	0.88	1.12
	B/C	2	0-18	194,746	0.79	1.20	105,627	0.73	1.16
4-5	A	2	0-18	60,923	0.89	1.04	19,218	0.87	1.06
	B/C	2	0-18	190,623	0.70	1.22	85,316	0.69	1.24
6-8	A	2	0-18	126,696	0.88	1.08	34,918	0.86	1.08
	B/C	2	0-18	174,202	0.70	1.03	69,451	0.70	1.03
9-12	A	2	0-18	128,368	0.89	1.04	31,543	0.87	1.07
	B/C	2	0-18	189,947	0.66	1.23	80,513	0.65	1.28

Table 5.1.3.4

Reliabilities of Domain Scores: Writ S601 Online by IEP Status

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	A	19,604	2	0-18	0.87	1.09
	B/C	897	2	0-18	0.81	1.06
2-3	A	21,327	2	0-18	0.87	1.15
	B/C	23,632	2	0-18	0.83	1.24
4-5	A	17,763	2	0-18	0.86	1.07
	B/C	27,642	2	0-18	0.75	1.25
6-8	A	34,224	2	0-18	0.82	1.09
	B/C	28,508	2	0-18	0.75	1.07
9-12	A	24,688	2	0-18	0.86	1.03
	B/C	35,058	2	0-18	0.64	1.24

5.1.4 Speaking

Table 5.1.4.1

Reliabilities of Domain Scores: Spek S601 Online

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	Pre-A	13,774	3	0-6	0.86	0.80
	A	101,779	6	0-18	0.84	1.33
	B/C	95,004	6	0-24	0.84	1.57
2-3	Pre-A	20,849	3	0-6	0.87	0.73
	A	122,067	6	0-18	0.84	1.38
	B/C	270,074	6	0-24	0.84	1.60
4-5	Pre-A	8,541	3	0-6	0.86	0.79
	A	57,741	6	0-18	0.85	1.33
	B/C	295,502	6	0-24	0.84	1.58
6-8	Pre-A	16,498	3	0-6	0.86	0.75
	A	96,708	6	0-18	0.84	1.32
	B/C	297,477	6	0-24	0.85	1.52
9-12	Pre-A	33,907	3	0-6	0.86	0.71
	A	189,736	6	0-18	0.85	1.36
	B/C	217,743	6	0-24	0.88	1.47

Table 5.1.4.2

Reliabilities of Domain Scores: Spek S601 Online by Gender

				Female			Male		
Cluster	Tier	No. of Tasks	Total Possible Raw Score Points	No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	Pre-A	3	0-6	5,382	0.87	0.78	6,522	0.86	0.81
	A	6	0-18	40,622	0.84	1.31	47,492	0.84	1.33
	B/C	6	0-24	41,391	0.84	1.57	40,741	0.84	1.57
2-3	Pre-A	3	0-6	7,827	0.87	0.72	10,235	0.86	0.75
	A	6	0-18	47,784	0.84	1.37	57,296	0.84	1.39
	B/C	6	0-24	114,283	0.84	1.60	119,037	0.84	1.61
4-5	Pre-A	3	0-6	3,297	0.85	0.79	4,105	0.86	0.78
	A	6	0-18	21,460	0.85	1.33	28,144	0.85	1.33
	B/C	6	0-24	117,829	0.83	1.59	134,698	0.84	1.58
6-8	Pre-A	3	0-6	6,599	0.85	0.73	7,602	0.86	0.75
	A	6	0-18	35,506	0.84	1.34	46,920	0.84	1.31
	B/C	6	0-24	114,677	0.86	1.53	139,478	0.85	1.51
9-12	Pre-A	3	0-6	12,381	0.84	0.70	16,769	0.87	0.70
	A	6	0-18	69,043	0.83	1.38	93,567	0.85	1.36
	B/C	6	0-24	85,252	0.87	1.48	100,765	0.88	1.47

Table 5.1.4.3

Reliabilities of Domain Scores: Spek S601 Online by Ethnicity

Cluster	Tier	No. of Tasks	Total Possible Raw Score Points	Hispanic			Other		
				No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	Pre-A	3	0-6	10,105	0.86	0.80	3,037	0.86	0.76
	A	6	0-18	72,463	0.84	1.33	26,242	0.82	1.33
	B/C	6	0-24	57,334	0.84	1.57	35,118	0.83	1.58
2-3	Pre-A	3	0-6	14,803	0.87	0.75	4,992	0.86	0.65
	A	6	0-18	89,084	0.84	1.38	28,912	0.81	1.38
	B/C	6	0-24	169,845	0.84	1.60	93,174	0.83	1.61
4-5	Pre-A	3	0-6	5,716	0.85	0.80	1,651	0.84	0.70
	A	6	0-18	40,579	0.85	1.34	12,991	0.81	1.32
	B/C	6	0-24	198,128	0.84	1.58	86,784	0.82	1.60
6-8	Pre-A	3	0-6	12,008	0.86	0.74	2,423	0.85	0.66
	A	6	0-18	70,192	0.84	1.32	19,642	0.80	1.31
	B/C	6	0-24	206,747	0.86	1.51	78,027	0.84	1.54
9-12	Pre-A	3	0-6	26,433	0.85	0.71	4,273	0.88	0.60
	A	6	0-18	138,369	0.85	1.36	39,702	0.81	1.37
	B/C	6	0-24	143,757	0.88	1.47	64,134	0.85	1.48

Table 5.1.4.4

Reliabilities of Domain Scores: Spek S601 Online by IEP Status

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	Pre-A	2,117	3	0-6	0.86	0.80
	A	11,395	6	0-18	0.85	1.35
	B/C	5,349	6	0-24	0.83	1.58
2-3	Pre-A	3,180	3	0-6	0.84	0.66
	A	19,151	6	0-18	0.81	1.37
	B/C	19,245	6	0-24	0.85	1.60
4-5	Pre-A	773	3	0-6	0.85	0.67
	A	11,955	6	0-18	0.79	1.33
	B/C	31,242	6	0-24	0.84	1.60
6-8	Pre-A	1,303	3	0-6	0.89	0.59
	A	19,439	6	0-18	0.81	1.30
	B/C	39,405	6	0-24	0.85	1.51
9-12	Pre-A	2,568	3	0-6	0.91	0.64
	A	30,506	6	0-18	0.85	1.34
	B/C	24,908	6	0-24	0.88	1.47

5.2 Interrater Agreement Rates

DRC raters score students' responses to the tasks included on the ACCESS Writing and Speaking tests. We describe the scoring of students' responses to these performance tasks in Section 3.2.2. DRC selects a sample of 20% of all responses scored, chosen at random during the operational scoring process, for double scoring. The tables in this section provide information on the interrater agreement rates that the DRC raters achieved. These tables show, for each of the tasks, the percentage of agreement between two raters who independently scored students' responses to that task.

For Writing, the first column in the tables shows the task, and the second column shows the number of responses that raters double scored. The next two columns show the percentages of agreement (%AG) and adjacent agreement (%AD) that the raters achieved. The last column shows the percentage of nonadjacent scores (%NA) that the raters assigned.

The Writing Scoring Scale defines six levels of performance ranging from 0 to 6, with the possibility of awarding a "plus" score between levels (e.g., 3, 3+, or 4 are all valid scores). We considered scores that matched or were contiguous as signifying **agreement** (%AG)—for example, if Rater 1 assigned a score of 3+ while Rater 2 assigned a score of 3, 3+, or 4. We considered scores that were one whole score point apart as **adjacent scores** (%AD)—for example, if Rater 1 assigned a score of 3+ while Rater 2 assigned a score of 2+ or 4+. Finally, if two raters assigned scores that were more than one whole score point apart, we considered those scores to be **nonadjacent scores** (%NA). Note that for Writing, DRC reports separate rates of interrater agreement for the raters' scoring of students' keyboarded responses and for the raters' scoring of students' handwritten responses.

For Speaking, the first column in the tables shows the task, and the second column shows the number of responses that raters double scored. The next two columns show the percentages of exact agreement (%EX) and adjacent score agreement (%AD) that the raters achieved. The last column shows the percentage of nonadjacent scores (%NA) that the raters assigned.

The Speaking Scoring Scale defines four levels of performance, ranging from 0 to 4. We considered scores that matched as demonstrating **exact agreement** (%EX). If the scores that two raters assigned differed by one level, we considered those scores to be **adjacent scores** (%AD).

Finally, if two raters assigned scores that were more than one level apart, we considered those scores to be **nonadjacent scores** (%NA). Note that the Speaking tasks that target PL 1—the three tasks in the Tier Pre-A forms and the first three tasks in the Tier A forms—are designed for beginning students and use a restricted subset of levels in the Speaking Scoring Scale, with only three possible score levels (see Part 1, Sections 2.1.4 and 3.2.4 for more detail). As the range of possible score levels is smaller for these tasks, the rater agreement rates tend to be higher. Therefore, it is not appropriate to compare the interrater agreement rates across tiers, especially when the tasks and the raw score ranges for the tasks being compared are different.

WIDA stipulates a minimum interrater agreement rate of 70%. For Writing, DRC defines "**agreement**" as being scored as adjacent agreement (AG). See Section 3.2.2 for more detail about how WIDA and DRC used the agreement rates to ensure that DRC maintains sufficient quality control throughout the course of scoring.

For Writing, the lowest interrater agreement rate was 94%. For Speaking, the lowest interrater agreement rate was 75%.

5.2.1 Listening

Interrater Agreement is not relevant for the domain of Listening, as all items are multiple choice items.

5.2.2 Reading

Interrater Agreement is not relevant for the domain of Listening, as all items are multiple choice items.

5.2.3 Writing

5.2.3.1 Grade 1

Table 5.2.3.1.1

Interrater Agreement: Writ 1 A S601 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	130,890	96	4	0
	2	130,584	97	3	0

Table 5.2.3.1.2

Interrater Agreement: Writ 1 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	9,578	98	2	0
	2	9,358	98	2	0

5.2.3.2 Grade 2–3

Table 5.2.3.2.1

Interrater Agreement: Writ 2-3 A S601 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	81,344	97	3	0
	2	93,638	97	3	0

Table 5.2.3.2.2

Interrater Agreement: Writ 2-3 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	133,046	95	5	0
	2	136,190	94	6	0

5.2.3.3 Grades 4–5

Table 5.2.3.3.1

Interrater Agreement: Writ 4-5 A S601 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	6,332	98	2	0
		KB	32,592	97	3	0
	2	HW	6,238	98	2	0
		KB	32,496	97	3	0

Table 5.2.3.3.2

Interrater Agreement: Writ 4-5 B/C S601 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	12,714	97	3	0
		KB	113,292	97	3	0
	2	HW	11,246	97	3	0
		KB	118,864	97	3	0

5.2.3.4 Grades 6–8

Table 5.2.3.4.1

Interrater Agreement: Writ 6-8 A S601 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	248	98	2	0
		KB	72,908	96	4	0
	2	HW	226	98	2	0
		KB	72,064	97	3	0

Table 5.2.3.4.2

Interrater Agreement: Writ 6-8 B/C S601 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	296	99	1	0
		KB	112,548	99	1	0
	2	HW	292	99	1	0
		KB	113,678	98	2	0

5.2.3.5 Grades 9–12

Table 5.2.3.5.1

Interrater Agreement: Writ 9-12 A S601 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	136	99	1	0
		KB	72,680	98	2	0
	2	HW	134	97	3	0
		KB	72,746	97	3	0

Table 5.2.3.5.2

Interrater Agreement: Writ 9-12 B/C S601 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	84	100	0	0
		KB	123,054	99	1	0
	2	HW	94	100	0	0
		KB	131,764	98	2	0

5.2.4 Speaking

5.2.4.1 Grade 1

Table 5.2.4.1.1

Interrater Agreement: Spek 1 Pre-A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	12,066	98	2	0
	2	11,236	98	2	0
	3	10,796	98	2	0

Table 5.2.4.1.2

Interrater Agreement: Spek 1 A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	60,502	99	1	0
	2	60,500	89	11	0
	3	60,286	98	2	0
	4	60,276	84	15	0
	5	59,136	99	1	0
	6	59,142	87	13	0

Table 5.2.4.1.3

Interrater Agreement: Spek 1 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	48,102	84	16	0
	2	48,102	85	15	0
	3	50,442	79	21	0
	4	50,438	79	21	0
	5	49,780	83	17	0
	6	49,782	80	20	0

5.2.4.2 Grade 2–3

Table 5.2.4.2.1

Interrater Agreement: Spek 2-3 Pre-A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	12,694	99	1	0
	2	14,064	98	2	0
	3	13,074	99	1	0

Table 5.2.4.2.2

Interrater Agreement: Spek 2-3 A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	68,896	99	1	0
	2	68,896	82	17	1
	3	71,272	99	1	0
	4	71,262	83	16	1
	5	69,694	99	1	0
	6	69,694	80	19	0

Table 5.2.4.2.3

Interrater Agreement: Spek 2-3 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	136,138	79	21	0
	2	136,136	76	23	1
	3	137,880	75	24	1
	4	137,880	76	24	1
	5	133,958	75	24	0
	6	133,958	76	23	1

5.2.4.3 Grades 4–5

Table 5.2.4.3.1

Interrater Agreement: Spek 4-5 Pre-A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	6,554	97	3	0
	2	6,280	98	2	0
	3	5,730	98	2	0

Table 5.2.4.3.2

Interrater Agreement: Spek 4-5 A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	33,860	98	2	0
	2	33,862	87	13	0
	3	33,956	99	1	0
	4	33,956	87	13	0
	5	32,806	99	1	0
	6	32,806	88	12	0

Table 5.2.4.3.3

Interrater Agreement: Spek 4-5 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	143,100	80	20	0
	2	143,098	76	24	0
	3	145,904	76	24	0
	4	145,904	75	25	0
	5	141,890	78	22	0
	6	141,892	78	22	0

5.2.4.4 Grades 6–8

Table 5.2.4.4.1

Interrater Agreement: Spek 6-8 Pre-A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	11,866	98	2	0
	2	12,026	98	2	0
	3	11,034	98	2	0

Table 5.2.4.4.2

Interrater Agreement: Spek 6-8 A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	55,656	99	1	0
	2	55,656	87	13	0
	3	58,332	99	1	0
	4	58,332	87	13	0
	5	55,240	99	1	0
	6	55,238	88	12	0

Table 5.2.4.4.3

Interrater Agreement: Spek 6-8 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	163,524	81	19	0
	2	163,524	82	18	0
	3	159,246	80	20	0
	4	159,246	79	21	0
	5	157,054	83	17	0
	6	157,044	81	19	0

5.2.4.5 Grades 9–12

Table 5.2.4.5.1

Interrater Agreement: Spek 9-12 Pre-A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	22,766	98	2	0
	2	22,032	97	3	0
	3	23,404	98	2	0

Table 5.2.4.5.2

Interrater Agreement: Spek 9-12 A S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	109,814	99	1	0
	2	109,820	84	15	1
	3	108,758	99	1	0
	4	108,756	84	16	1
	5	112,070	99	1	0
	6	112,074	82	17	1

Table 5.2.4.5.3

Interrater Agreement: Spek 9-12 B/C S601 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	120,230	78	21	1
	2	120,222	81	19	0
	3	117,282	77	22	1
	4	117,280	76	23	1
	5	122,766	75	24	1
	6	122,766	76	23	2

5.3 Conditional Standard Errors of Measurement of the Scale Scores at the Cut Points

The tables in this section present information about the conditional standard errors of measurement (CSEM) values of scale scores at the most important points at which policy makers make decisions such as reclassification about students based on performance on ACCESS—the cut points between language proficiency levels. The CSEM provides information about the amount of measurement error we would expect in any student's scale score at that point on the underlying latent ability scale. We first computed CSEM values on the theta metric, which is the square root of the inverse of the Test Information Function. Next, we used the multiplicative constant of the linear equation for the domain to linearly transform those logit-based CSEM values so that we could report them on the ACCESS score scale (See Section 2.2).

When calculated using an IRT approach, CSEM values can vary across the scale scores. For example, in the Listening and Reading domains, if a student answers correctly either a very few or a very large number of items (i.e., scores at the extremes of the scale score distribution), the CSEM value will be larger than it would be if the student correctly answers a moderate number of items. Scale scores near the middle of the score distribution typically have lower CSEM values compared to scale scores near the extremes because many tests are comprised of a large proportion of moderately difficult items, which are well suited to measuring students of moderate proficiency.

We use the CSEM to construct an error band, quantifying the amount of uncertainty in a student's scale score. One CSEM below a student's scale score and one CSEM above that scale score indicates an approximate 68% confidence interval. To interpret this confidence interval, consider a student who takes the test 100 times. Assuming measurement error is normally distributed, the student's true proficiency would fall within the confidence interval 68% of the time (or 68 times out of 100).

As a rule, lower CSEM values around scale scores at important decision points are desirable. Generally speaking, the most important decision points for the ACCESS scores are at the PL 3/4 and PL 4/5 cut points, although the approaches that WIDA states use to make decisions about ACCESS scores differ. As discussed in Section 5, all WIDA states use composite scale scores when making reclassification decisions, and no WIDA state uses a single domain scale score

when making those decisions. Because each grade has its own set of cut points, we provide information for each grade within a grade-level cluster.

Since we scale ACCESS test scores using an IRT approach, CSEM values for the scale scores at the highest cut points are typically large. Use of this approach tends to produce larger CSEM values at the lower and the higher ends of the score scale. In addition, because students exit the EL program when they demonstrate that they are English language proficient, there are typically fewer students at the highest cut points than at those other cut points. Therefore, the CSEM values associated with the scale scores at the highest cut points tend to be larger than those of the scale scores at the lower cut points since there are fewer students available for estimating the scores and the CSEM values for these scores.

Since the Listening and Reading tests are multistage adaptive tests, the CSEM values will vary for the same scale score because the test will route students to take different items; therefore, it is not possible to present a single CSEM value for the scale score that corresponds to each cut point. In the tables for Listening and Reading, the leftmost column shows the proficiency level cut (e.g., 1/2, which is the cut between PL 1 and PL 2). The second column shows the grade level. The third column shows the cut point in the scale score metric (e.g., 305). The next columns present the number of students and the minimum, maximum, mean, and standard deviation of the CSEM values for all students' scale scores at each cut point within a grade level. Note that there are some rare cases where there are no observed scale scores corresponding to certain cut points; therefore, we cannot provide these descriptive statistics. Because Listening and Reading tests are multistage adaptive tests, we would not expect large variation in the mean CSEM values of students' scale scores across cut points within a grade level.

For Writing and Speaking, we present the CSEM values for the scale scores by tier. From these tables, it is possible to determine the extent to which students' responses to the tasks included in the different Writing and Speaking tiers provide targeted information that is useful for accurately placing them into the various proficiency levels. In the tables for Writing and Speaking, the leftmost column shows the proficiency level cut point (e.g., 1/2, which is the cut between PL 1 and PL 2). The second column shows the grade level. The third column shows the cut point in the scale score metric (e.g., 305). In the last column(s), the corresponding CSEM value for the scale score at each cut point are shown.

5.3.1 Listening

5.3.1.1 Grade 1

Table 5.3.1.1

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: List 1 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	1	236	859	18.37	18.88	18.80	0.19
2/3	1	259	165	16.33	16.33	16.33	0.00
3/4	1	291	37	16.33	17.86	17.00	0.64
4/5	1	303	1,975	16.33	17.35	16.41	0.22
5/6	1	327	235	17.86	18.37	18.23	0.23

5.3.1.2 Grades 2-3

Table 5.3.1.2

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: List 2-3 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	2	245	N/A	N/A	N/A	N/A	N/A
	3	262	13	18.37	21.94	20.84	1.72
2/3	2	283	N/A	N/A	N/A	N/A	N/A
	3	300	505	19.39	19.90	19.84	0.16
3/4	2	314	2,830	18.37	21.94	18.38	0.20
	3	331	417	17.86	19.90	18.23	0.75
4/5	2	330	2,266	18.37	20.41	20.18	0.63
	3	349	1,702	17.86	26.02	18.50	1.37
5/6	2	354	234	19.90	23.47	19.94	0.40
	3	374	113	22.96	22.96	22.96	0.00

5.3.1.3 Grades 4-5

Table 5.3.1.3

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: List 4-5 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	4	275	3	17.35	17.35	17.35	0.00
	5	285	71	17.35	19.39	19.19	0.59
2/3	4	313	133	16.33	17.86	16.71	0.44
	5	323	420	15.82	16.84	16.08	0.26
3/4	4	343	1,767	16.33	17.86	16.33	0.08
	5	354	1,329	17.35	17.86	17.35	0.05
4/5	4	363	143	17.35	17.86	17.83	0.12
	5	375	31	17.86	18.37	18.29	0.19
5/6	4	388	5	18.37	18.37	18.37	0.00
	5	401	116	19.39	19.39	19.39	0.00

5.3.1.4 Grades 6-8

Table 5.3.1.4

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: List 6-8 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	6	294	300	19.39	20.41	20.31	0.30
	7	302	N/A	N/A	N/A	N/A	N/A
	8	308	32	19.90	19.90	19.90	0.00
2/3	6	332	11	15.82	15.82	15.82	0.00
	7	340	106	15.82	16.33	16.03	0.25
	8	347	223	15.82	15.82	15.82	0.00
3/4	6	363	9	16.33	16.33	16.33	0.00
	7	370	3	15.82	15.82	15.82	0.00
	8	377	30	15.82	16.84	16.77	0.22
4/5	6	385	1,516	16.84	17.35	16.90	0.17
	7	394	464	16.84	16.84	16.84	0.00
	8	402	176	17.86	18.88	18.00	0.35
5/6	6	411	13	17.86	17.86	17.86	0.00
	7	420	N/A	N/A	N/A	N/A	N/A
	8	427	N/A	N/A	N/A	N/A	N/A

5.3.1.5 Grades 9-12

Table 5.3.1.5

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: List 9-12 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	9	314	3,915	19.39	19.39	19.39	0.00
	10	325	2,806	18.88	19.39	18.88	0.04
	11	335	10	18.88	18.88	18.88	0.00
	12	342	138	16.84	19.90	19.61	0.49
2/3	9	353	37	16.33	18.37	16.38	0.34
	10	358	35	16.33	19.39	16.41	0.52
	11	364	2	16.33	16.33	16.33	0.00
	12	368	23	16.33	16.33	16.33	0.00
3/4	9	383	49	16.84	16.84	16.84	0.00
	10	389	506	16.84	16.84	16.84	0.00
	11	394	845	16.84	17.35	16.84	0.05
	12	398	478	16.84	16.84	16.84	0.00
4/5	9	409	179	16.84	17.35	17.22	0.22
	10	415	65	16.84	17.35	16.94	0.21
	11	420	210	17.35	19.39	17.96	0.29
	12	426	62	17.35	18.88	18.33	0.65
5/6	9	434	19	18.88	18.88	18.88	0.00
	10	441	56	17.86	20.41	18.90	1.27
	11	447	N/A	N/A	N/A	N/A	N/A
	12	452	2,696	18.37	19.90	18.43	0.31

5.3.2 Reading

5.3.2.1 Grade 1

Table 5.3.2.1

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Read 1 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	1	264	2,105	10.71	12.76	12.03	0.40
2/3	1	286	1,708	9.69	10.71	9.87	0.24
3/4	1	304	509	9.69	10.20	9.98	0.25
4/5	1	315	65	9.69	10.20	9.79	0.20
5/6	1	334	2,080	10.20	10.71	10.22	0.08

5.3.2.2 Grades 2-3

Table 5.3.2.2

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Read 2-3 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	2	283	350	11.22	12.76	11.69	0.54
	3	297	157	10.20	10.71	10.41	0.25
2/3	2	307	527	10.20	10.71	10.28	0.19
	3	323	2,027	9.69	10.20	9.70	0.03
3/4	2	326	8,609	9.69	10.20	10.19	0.09
	3	342	6,942	9.69	10.20	9.71	0.10
4/5	2	337	166	9.69	10.20	9.71	0.08
	3	352	58	10.20	10.71	10.30	0.20
5/6	2	355	8	10.20	10.20	10.20	0.00
	3	370	21	11.73	11.73	11.73	0.00

5.3.2.3 Grades 4-5

Table 5.3.2.3

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Read 4-5 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	4	307	1,798	10.71	12.24	11.79	0.29
	5	316	859	10.20	12.24	10.97	0.36
2/3	4	335	4,899	10.20	11.22	10.29	0.19
	5	345	786	9.69	10.71	10.20	0.08
3/4	4	354	1,004	9.69	10.71	10.20	0.09
	5	364	1,260	10.20	10.71	10.26	0.16
4/5	4	364	540	10.20	10.71	10.24	0.12
	5	373	400	10.20	10.71	10.53	0.24
5/6	4	382	15	10.71	10.71	10.71	0.00
	5	391	6	11.73	11.73	11.73	0.00

5.3.2.4 Grades 6-8

Table 5.3.2.4

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Read 6-8 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	6	323	808	11.73	12.76	12.45	0.26
	7	329	796	11.73	12.76	12.22	0.22
	8	335	298	11.22	12.24	11.82	0.28
2/3	6	353	582	10.20	10.71	10.31	0.21
	7	360	2,015	10.20	11.22	10.24	0.15
	8	366	2,801	10.20	11.22	10.24	0.13
3/4	6	373	1,027	10.20	10.71	10.27	0.17
	7	380	981	10.20	11.22	10.32	0.22
	8	386	1,308	10.20	12.24	10.55	0.35
4/5	6	382	771	10.20	11.22	10.30	0.20
	7	389	2,253	10.20	11.22	10.23	0.12
	8	395	504	10.71	11.73	10.84	0.23
5/6	6	399	21	10.71	10.71	10.71	0.00
	7	406	14	10.71	12.24	11.01	0.56
	8	412	2,147	11.22	11.73	11.24	0.09

5.3.2.5 Grades 9-12

Table 5.3.2.5

Descriptive Statistics for the Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Read 9-12 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	9	340	4,950	11.22	12.24	11.26	0.19
	10	344	2,054	11.22	12.24	11.37	0.24
	11	348	248	11.22	12.76	11.28	0.23
	12	352	110	11.22	12.76	11.41	0.43
2/3	9	372	1,045	10.20	11.22	10.70	0.08
	10	377	458	10.20	11.22	10.73	0.09
	11	382	896	10.20	11.22	10.30	0.27
	12	386	215	10.20	10.71	10.28	0.18
3/4	9	392	2,816	10.20	11.22	10.44	0.25
	10	397	1,621	10.20	11.22	10.28	0.18
	11	402	436	10.20	11.73	10.35	0.24
	12	407	190	10.20	11.22	10.35	0.24
4/5	9	401	4,578	10.20	11.22	10.24	0.14
	10	406	484	10.20	11.22	10.57	0.23
	11	410	310	10.20	11.22	10.62	0.27
	12	414	146	10.20	11.73	10.69	0.39
5/6	9	418	3,148	10.20	12.24	10.21	0.08
	10	423	369	10.71	11.73	10.72	0.08
	11	427	190	11.22	12.24	11.27	0.16
	12	432	105	11.22	12.24	11.31	0.29

5.3.3 Writing

5.3.3.1 Grade 1

Table 5.3.3.1

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Writ 1 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	1	238	15.31	14.66
2/3	1	275	20.94	20.41
3/4	1	337	20.68	20.94
4/5	1	382	19.33	18.80
5/6	1	405	25.24	23.09

5.3.3.2 Grades 2-3

Table 5.3.3.2

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Writ 2-3 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	2	242	14.61	14.23
	3	247	15.31	14.23
2/3	2	279	20.41	19.33
	3	283	20.94	19.87
3/4	2	341	20.94	21.48
	3	346	20.68	21.21
4/5	2	388	19.06	18.53
	3	394	19.60	18.73
5/6	2	411	23.90	21.21
	3	418	26.82	23.09

5.3.3.3 Grades 4-5

Table 5.3.3.3

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Writ 4-5 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	4	266	14.24	22.56
	5	267	14.23	21.75
2/3	4	288	16.92	15.12
	5	293	17.92	14.77
3/4	4	351	21.75	21.21
	5	356	21.75	21.48
4/5	4	401	18.80	20.94
	5	407	18.53	20.68
5/6	4	425	19.60	19.33
	5	433	21.21	18.84

5.3.3.4 Grades 6-8

Table 5.3.3.4

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Writ 6-8 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	6	268	15.31	14.50
	7	273	16.11	14.50
	8	281	17.45	15.31
2/3	6	298	20.14	18.26
	7	305	20.94	19.33
	8	311	21.21	20.14
3/4	6	361	20.94	21.67
	7	367	20.68	21.48
	8	372	20.14	21.21
4/5	6	413	19.33	18.80
	7	419	20.41	18.80
	8	424	21.48	19.06
5/6	6	441	27.12	22.29
	7	450	31.95	25.24
	8	459	37.86	29.54

5.3.3.5 Grades 9-12

Table 5.3.3.5

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Writ 9-12 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	9	289	14.23	14.76
	10	298	14.50	15.00
	11	308	15.84	16.38
	12	318	17.72	17.99
2/3	9	319	17.94	18.26
	10	326	19.06	19.33
	11	335	20.41	20.41
	12	344	20.94	21.21
3/4	9	378	21.75	21.75
	10	385	21.75	21.48
	11	391	21.48	21.21
	12	398	20.94	20.74
4/5	9	430	18.72	18.80
	10	436	18.53	18.80
	11	441	18.80	19.06
	12	447	19.06	19.60
5/6	9	469	23.90	24.97
	10	479	27.93	29.00
	11	490	34.37	35.44
	12	501	42.16	43.50

5.3.4 Speaking

5.3.4.1 Grade 1

Table 5.3.4.1

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Spek 1 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	1	205	21.94	15.79
2/3	1	261	28.08	19.89
3/4	1	311	23.98	17.26
4/5	1	361	30.71	20.77
5/6	1	403	52.06	33.63

Note: Tier Pre-A is not presented as it is not possible for Tier Pre-A students to receive a proficiency level higher than 2.

5.3.4.2 Grades 2-3

Table 5.3.4.2

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Spek 2-3 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	2	220	22.52	15.79
	3	234	24.57	16.96
2/3	2	273	28.08	19.89
	3	283	27.49	19.60
3/4	2	322	24.28	17.26
	3	332	24.26	16.96
4/5	2	374	30.42	20.47
	3	386	34.51	23.11
5/6	2	415	50.60	33.34
	3	425	58.50	38.61

Note: Tier Pre-A is not presented as it is not possible for Tier Pre-A students to receive a proficiency level higher than 2.

5.3.4.3 Grades 4-5

Table 5.3.4.3

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Spek 4-5 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	4	246	21.94	15.79
	5	258	23.11	16.09
2/3	4	293	28.08	18.72
	5	302	28.66	19.30
3/4	4	342	24.86	18.43
	5	350	24.28	17.84
4/5	4	397	27.49	18.13
	5	407	30.13	19.30
5/6	4	435	42.12	25.15
	5	443	46.80	27.49

Note: Tier Pre-A is not presented as it is not possible for Tier Pre-A students to receive a proficiency level higher than 2.

5.3.4.4 Grades 6-8

Table 5.3.4.4

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Spek 6-8 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	6	268	22.81	15.79
	7	277	23.98	16.38
	8	284	25.22	16.96
2/3	6	310	28.37	19.60
	7	317	28.37	19.97
	8	323	28.08	20.18
3/4	6	360	24.28	17.84
	7	369	23.98	17.26
	8	377	23.98	17.10
4/5	6	417	29.54	19.60
	7	425	31.88	20.77
	8	433	35.10	22.52
5/6	6	451	44.46	27.79
	7	457	48.55	29.83
	8	463	52.94	32.46

Note: Tier Pre-A is not presented as it is not possible for Tier Pre-A students to receive a proficiency level higher than 2.

5.3.4.5 Grades 9-12

Table 5.3.4.5

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Spek 9-12 S601 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	9	290	26.03	17.26
	10	295	26.62	17.84
	11	299	27.20	18.43
	12	302	27.49	18.43
2/3	9	328	27.79	19.89
	10	333	27.20	19.60
	11	337	26.83	19.60
	12	340	26.62	19.30
3/4	9	385	24.28	17.26
	10	393	24.86	17.26
	11	400	25.76	17.63
	12	406	26.91	18.13
4/5	9	440	37.73	23.40
	10	446	40.95	25.15
	11	451	43.58	26.62
	12	455	46.21	28.08
5/6	9	468	56.16	33.34
	10	471	58.79	34.51
	11	474	61.13	35.97
	12	476	63.17	37.14

Note: Tier Pre-A is not presented as it is not possible for Tier Pre-A students to receive a proficiency level higher than 2.

5.4 Accuracy and Consistency of Domains

One of the main purposes of the WIDA ACCESS program is to identify the English language proficiency level of students with respect to the WIDA ELD Standards. Because of the emphasis on the classification of student performance, a question of interest is how accurately and consistently ACCESS domain scale scores can classify students into the WIDA proficiency levels determined by the 2016 ACCESS standard-setting process (Cook & MacGregor, 2017). Test users can examine indices that report on the accuracy and consistency of these classifications and can use that information to judge the utility of WIDA's proficiency level categorization, while policy makers can use these indices to assist them when making decisions about ACCESS test design and score reporting (American Educational Research Association et al., 2014). The analyses we conduct to examine the accuracy and consistency of classifications utilize the methods that Livingston and Lewis (1995) and Young and Yoon (1998) outlined, as implemented in the software program BB-CLASS (Brennan, 2004; cf. also Lee, Hanson, & Brennan, 2002).

Classification accuracy is defined conceptually as the extent to which the proficiency classifications of students based on their observed raw scores or scale scores would agree with those made based on their true scores (Livingston, 2018; Livingston & Lewis, 1995). A student's true score is the average of the scores that the student would have received, averaging over some set of prespecified factors or conditions (e., different versions of the test, different times of test administration). Therefore, the calculation of the true scores depends upon the particular factors over which one chooses to average (Livingston, 2018). We assume that true scores measure perfectly, but those scores are unknown. Therefore, to provide the best estimation of classification accuracy for WIDA, we use test data from one ACCESS administration to estimate students' true scale scores based on their domain scale scores and the parameters of the model used in estimating those true scale scores. We can then use the results from our analysis to estimate the percentages of the students who were accurately classified into each proficiency level.

Classification consistency is defined conceptually as the extent to which the proficiency classifications of students agree, given two independent administrations of the same or two parallel test forms. It is impractical to obtain repeated administrations of the same or parallel test

forms because of cost, testing burden, and the effects of student memory and practice. However, it is possible to estimate the percentages of the students who would be consistently classified with the assumption that the same test is independently administered twice to the same group of students.

The approach that Livingston and Lewis (1995) took, which we implemented here, uses information about the reliability of the students' domain scale scores, the cut points, and the observed distribution of scores. Then, using a four-parameter beta distribution, we model the distribution of the true scale scores and of the domain scale scores on a parallel form. The Livingston and Lewis procedure requires that the reliability estimate of the students' scores on a test form be provided when calculating the classification consistency and accuracy indices. For Listening and Reading, we used the Rasch student separation reliability estimates by grade-level clusters in the procedure. Since the Writing and Speaking tests were tiered, we needed to produce a single reliability estimate across tiers to implement the Livingston and Lewis procedure. This is a weighted reliability estimate across tiers (see Section 5.1).

Overall Classification Accuracy and Consistency

Overall classification accuracy indicates the percentage of all students whom we would classify into the same language proficiency level by both their domain scale scores and their true scale scores (i.e., the percentage of students whom we accurately classified). For example, an overall classification accuracy index of 0.774 means that we would classify 77% of the students into the same proficiency level according to their domain scale scores and their true scale scores. **Overall classification consistency** indicates the percentage of all students whom we would classify into the same language proficiency levels by their performances on both the administered test and on a parallel test. For example, an overall classification consistency index of 0.664 means that we would classify 66% of the students into the same proficiency level if they took two parallel forms of the test. A classification consistency index is always lower than its corresponding classification accuracy index, because in classification consistency, a classification based on a student's performance on the administered test and a classification based on that student's performance on a parallel test are both subject to measurement error. In contrast, in classification accuracy, only the classification based on a student's performance on the administered test

contains error while we assume that the classification based on that student's true scale score is free of measurement error.

Marginal Classification Accuracy and Consistency

Overall classification accuracy and consistency indices indicate the degree to which we accurately and consistently classify students into the same WIDA proficiency levels, but not the degree to which we accurately or consistently classify students into the proficiency levels below or above the specific cut point (e.g., at the PL 4/PL 5 cut point). The indices that can address this question are **marginal classification accuracy and consistency indices based on domain scale scores at the cut points**. From an accountability perspective, the most important indices for test users and policy makers to examine are the marginal classification accuracy and consistency indices.

The **marginal classification accuracy indices based on domain scale scores at the cut points** report the percentage of students whom we accurately placed into proficiency levels above and below each cut point based on their domain scale scores. For example, a classification accuracy index of 0.774 at the PL 4/PL 5 cut point means that we would classify 77% of the students in the same way using their domain scale scores or their true scale scores, either into the proficiency levels below the cut point (i.e., PL 1 to PL 4) or into the proficiency levels above the cut point (i.e., PL 5 to PL 6). The **marginal classification consistency indices based on domain scale scores at the cut points** report the percentage of students whom we would classify consistently above and below each cut point based on their domain scale scores. For example, a classification consistency index of 0.664 at the PL 4/PL 5 cut point means that we would classify 66% of the students in the same way if they took two parallel forms, either into the proficiency levels below the cut point (i.e., PL 1 to PL 4) or into the proficiency levels above the cut point (i.e., PL 5 to PL 6). Note that the marginal accuracy and consistency indices are generally higher for students' domain scale scores at the cut points than are the overall classification accuracy and consistency indices (Livingston, 2018). This is because the marginal accuracy and consistency indices report the classification decisions at one cut point at a time while the overall accuracy and consistency indices report the classification decisions at all five cut points at the same time.

The interactions of a number of factors affect the calculation of classification accuracy and consistency: (1) the number of proficiency level cut points, (2) the magnitude of the test score

reliability coefficient, (3) measurement accuracy for scale scores at the cut points, (4) the distances between adjacent cut points, (5) the locations of the cut points on the ability scale, and (6) the proportion of students' scale scores around a cut point (Ercikan & Julian, 2002; Lee et al., 2002). These factors are functions of the test design and, most importantly, the standard-setting decisions. The indices are lower when there is a greater number of proficiency levels, a lower test score reliability coefficient, and higher measurement accuracy of the scale scores at the cut points, as well as when the two adjacent cut points are closer, and when more students' domain scale scores are around a cut point. Furthermore, the numbers and types of items on a test affect the calculation of the test score reliability coefficient. The lower the test score reliability, the lower the classification accuracy and consistency indices would be. For example, the test score reliability coefficient for the ACCESS Online Writing domain raw scores would be lower than the test score reliability coefficients for similar tests that include more items or tasks since we estimate the test score reliability coefficient for ACCESS Online Writing domain raw scores based on students' performance on only two tasks. Therefore, the classification accuracy and consistency indices for the Writing domain might be lower than those for other domains.

For each test domain, we present three tables. The first reports indices that describe the overall accuracy and overall consistency of the proficiency level classifications for each grade level. The second reports the marginal classification accuracy indices based on domain scale scores at the cut points for each grade level. The third reports the marginal classification consistency indices based on domain scale scores at the cut points for each grade level. If we could not estimate the overall and marginal classification accuracy and consistency indices because we classified fewer than 200 students into a given proficiency level, we combined the affected proficiency level and the proficiency level below it and placed 'N/A' in the table for the affected proficiency level.

Assessment experts have issued little guidance to aid in making judgments about the ideal or expected levels of decision consistency and accuracy needed for educational assessments since many different factors affect the calculation of these indices, as discussed earlier. To help test users and policy makers interpret the results from our classification analyses, for each of the ACCESS test domains, we report the range of the overall classification accuracy and consistency indices across grades. Additionally, we highlight the grade with the lowest classification accuracy and consistency indices. Since the overall accuracy and consistency indices are summaries of the degree of classification accuracy and consistency across all proficiency level

cut points, we also report the marginal classification accuracy and consistency indices for these grades to identify the specific source(s) of low classification accuracy and consistency.

For Listening, as shown in Table 5.4.1.1, the overall classification accuracy indices ranged from 0.574 to 0.772, and the overall classification consistency indices ranged from 0.466 to 0.714. Grade 11 had the lowest overall classification accuracy and consistency indices for Listening.

For Reading, as shown in Table 5.4.2.1, the overall classification accuracy indices ranged from 0.597 to 0.688, and the overall classification consistency indices ranged from 0.486 to 0.592. Grade 1 had the lowest overall classification accuracy and consistency indices for Reading.

For Writing, as shown in Table 5.4.3.1, the overall classification accuracy indices ranged from 0.565 to 0.782, and the overall classification consistency indices ranged from 0.499 to 0.677. Grade 4 had the lowest overall classification accuracy and consistency indices for Writing.

For Speaking, as shown in Table 5.4.4.1, the overall classification accuracy indices ranged from 0.624 to 0.751, and the overall classification consistency indices ranged from 0.522 to 0.656. Grade 5 had the lowest overall classification accuracy indices for Speaking, while Grade 5 had the lowest overall classification consistency index.

From an accountability perspective, the most important indices for test users and policy makers to examine are the marginal classification accuracy and consistency indices. To help them interpret our results, we report for each domain the range of the marginal classification accuracy and consistency indices across grades and then highlight the grades (and the cut points within those grades) that had the lowest marginal classification accuracy and the lowest classification consistency.

For Listening, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.874 to 0.987 (Table 5.4.1.2), and the marginal classification consistency indices ranged from 0.825 to 0.982 (Table 5.4.1.3). Grade 8, at the PL 4/5 cut point, had the lowest marginal classification accuracy and consistency indices.

For Reading, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.831 to 0.982 (Table 5.4.2.2), and the marginal classification consistency indices ranged from 0.774 to 0.972 (Table 5.4.2.3). Grade 1, at the PL 1/2 cut point, had the lowest marginal classification accuracy and consistency indices. Note that Grade 1 also had the lowest

overall classification accuracy index in the Reading domain. The low marginal classification accuracy and consistency at the PL 1/2 cut point appeared to have contributed to its low overall classification accuracy. However, it should be noted that the marginal classification accuracy and consistency indices for Grade 1 Reading are still in the 0.70 to 0.90 range.

For Writing, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.657 to 0.998 (Table 5.4.3.2), and the marginal classification consistency indices ranged from 0.622 to 0.998 (Table 5.4.3.3). Grade 4, at the PL 3/4 cut point, had the lowest marginal classification accuracy and consistency indices. Note that Grade 4 also had the lowest overall classification accuracy and consistency indices in the Writing domain. For Grade 4, the low marginal classification accuracy and consistency at the PL 3/4 cut point appeared to have contributed to their low overall classification accuracy and consistency.

For Speaking, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.814 to 0.998 (Table 5.4.4.2), and the marginal classification consistency indices ranged from 0.750 to 0.998 (Table 5.4.4.3). Grade 5, at the PL 3/4 cut point, had the lowest marginal classification accuracy and consistency indices. However, it should be noted that the marginal classification accuracy and consistency indices for Grade 5 Speaking are still in the 0.70 to 0.90 range.

When we compared the overall and marginal classification accuracy and consistency indices based on the domain scale scores for a particular grade, we saw that in many instances they told the same story (i.e., for a given grade, when the overall classification accuracy and consistency indices were low, then the marginal classification accuracy and consistency indices also tended to be low).

We observed that in the domains of Listening, Writing, and Speaking, the marginal classification accuracy and consistency indices for PL cut points in the middle of the proficiency level range (i.e., PL 3/4 cut points) tended, on average, to be lower than the marginal classification accuracy and consistency indices for cut points at the lower and upper ends of the range, a finding that is consistent with findings from previous researchers (Ercikan & Julian, 2002; Lee et al., 2002). One possible reason might be that the cut points for the proficiency levels in the middle of the proficiency level range tend to be closer together than the cut points for the proficiency levels at the ends of that range. (Cut points tend to be closer to each other when there are a large number

of proficiency levels.) We would expect marginal classification accuracy and consistency to vary for different ability levels due to variation in measurement accuracy. That is, the further away the students' domain scale scores are from the cut points, the smaller the classification errors would be, or the more accurate the classification decisions would be. With many proficiency levels, there are more student domain scale scores near the cut points than there would be if there were fewer proficiency levels. Therefore, the higher the number of proficiency levels, the higher the probability that we would misclassify students (Ercikan & Julian, 2002). Additionally, the intervals between cut points that are in the middle of the ACCESS proficiency level range are smaller than the intervals between cut points that are at the upper and lower ends of the proficiency level range. Consequently, the marginal classification accuracy and consistency indices based on the domain scale scores for the PL 2/3 and PL 3/4 cut points tend to be lower than for other cut points, as we might expect.

Although assessment experts have issued little guidance to aid in making judgments about the ideal or expected levels of decision consistency and accuracy needed for educational assessments since many different factors affect the calculation of these indices, as discussed earlier, the ranges of the classification accuracy and consistency indices for the ACCESS domains are very similar to those reported for similar testing programs such as ELPA21 (American Institutes of Research, 2018), with the exception of the Writing domain. Since the ACCESS Online Writing test consists of only two tasks, the test score reliability estimate may be lower than similar writing tests that include more tasks. The classification accuracy and consistency indices derived using the Livingston and Lewis (1995) procedure are affected by the magnitude of the test score reliability, which is lower when a test has fewer tasks. Also note that we would not expect the indices estimated for ACCESS domains to be exactly the same as those computed in other programs, because testing programs differ in their student populations, the numbers of proficiency levels, their test designs, their score distributions, and the methods used to compute classification accuracy and consistency indices. For example, compared to similar testing programs, students taking ACCESS represent a much larger and more diverse population. Additionally, the ACCESS testing program defines more proficiency levels than other similar testing programs, and the ACCESS test design is more complex. Therefore, it is difficult to compare the classification accuracy and consistency indices for ACCESS domains to those for other testing programs.

5.4.1 Listening

Table 5.4.1.1

Overall Accuracy and Consistency of Classification Indices: List S601 Online

Grade	Accuracy	Consistency
1	0.661	0.585
2	0.594	0.502
3	0.591	0.501
4	0.772	0.714
5	0.727	0.665
6	0.610	0.510
7	0.597	0.500
8	0.578	0.481
9	0.584	0.475
10	0.577	0.468
11	0.574	0.466
12	0.581	0.474

Table 5.4.1.2

Marginal Classification Accuracy Indices Based on the Domain Scale Scores at the Cut Points: List S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.941	0.932	0.911	0.907	0.899
2	0.946	0.920	0.887	0.882	0.900
3	0.948	0.916	0.883	0.885	0.899
4	0.987	0.969	0.946	0.936	0.898
5	0.976	0.960	0.938	0.921	0.886
6	0.983	0.940	0.899	0.875	0.884
7	0.974	0.930	0.892	0.878	0.889
8	0.962	0.922	0.887	0.874	0.892
9	0.954	0.902	0.879	0.895	0.930
10	0.945	0.900	0.874	0.898	0.935
11	0.935	0.905	0.876	0.892	0.939
12	0.927	0.898	0.880	0.907	0.943

Table 5.4.1.3

Marginal Classification Consistency Indices Based on the Domain Scale Scores at the Cut Points: List S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.918	0.902	0.875	0.869	0.859
2	0.925	0.884	0.843	0.837	0.860
3	0.926	0.879	0.840	0.839	0.858
4	0.982	0.957	0.925	0.904	0.856
5	0.967	0.942	0.911	0.884	0.841
6	0.975	0.916	0.857	0.826	0.839
7	0.962	0.901	0.849	0.830	0.846
8	0.947	0.888	0.842	0.825	0.848
9	0.934	0.862	0.832	0.853	0.901
10	0.922	0.858	0.826	0.855	0.907
11	0.909	0.863	0.829	0.850	0.911
12	0.898	0.856	0.834	0.867	0.919

5.4.2 Reading

Table 5.4.2.1

Overall Accuracy and Consistency of Classification Indices: Read S601 Online

Grade	Accuracy	Consistency
1	0.597	0.486
2	0.616	0.502
3	0.602	0.498
4	0.606	0.502
5	0.617	0.515
6	0.688	0.592
7	0.680	0.585
8	0.673	0.581
9	0.658	0.560
10	0.643	0.547
11	0.640	0.543
12	0.657	0.560

Table 5.4.2.2

Marginal Classification Accuracy Indices Based on the Domain Scale Scores at the Cut Points: Read S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.831	0.886	0.925	0.955	0.978
2	0.939	0.891	0.893	0.911	0.962
3	0.902	0.897	0.903	0.912	0.950
4	0.948	0.908	0.880	0.894	0.946
5	0.936	0.896	0.888	0.907	0.953
6	0.915	0.900	0.924	0.949	0.982
7	0.913	0.903	0.924	0.943	0.973
8	0.913	0.906	0.919	0.940	0.967
9	0.925	0.907	0.914	0.925	0.956
10	0.926	0.905	0.903	0.918	0.956
11	0.928	0.901	0.900	0.915	0.953
12	0.921	0.902	0.907	0.925	0.962

Table 5.4.2.3

Marginal Classification Consistency Indices Based on the Domain Scale Scores at the Cut Points: Read S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.774	0.838	0.898	0.934	0.968
2	0.913	0.846	0.852	0.878	0.944
3	0.864	0.852	0.867	0.881	0.928
4	0.927	0.867	0.839	0.854	0.921
5	0.910	0.853	0.847	0.870	0.931
6	0.882	0.861	0.894	0.925	0.972
7	0.878	0.866	0.893	0.918	0.960
8	0.877	0.870	0.889	0.913	0.953
9	0.897	0.869	0.880	0.895	0.936
10	0.897	0.865	0.868	0.887	0.935
11	0.901	0.862	0.864	0.881	0.930
12	0.891	0.863	0.875	0.894	0.943

5.4.3 Writing

Table 5.4.3.1

Overall Accuracy and Consistency of Classification Indices: Writ S601 Online

Grade	Accuracy	Consistency
1	0.672	0.610
2	0.729	0.633
3	0.733	0.634
4	0.565	0.499
5	0.651	0.521
6	0.702	0.620
7	0.782	0.677
8	0.734	0.631
9	0.637	0.548
10	0.704	0.586
11	0.654	0.550
12	0.660	0.570

Table 5.4.3.2

Marginal Classification Accuracy Indices Based on the Domain Scale Scores at the Cut Points: Writ S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.883	0.788	0.989	N/A	N/A
2	0.934	0.831	0.960	N/A	N/A
3	0.953	0.902	0.875	0.998	N/A
4	0.967	0.930	0.657	0.971	0.998
5	0.967	0.938	0.737	0.984	0.996
6	0.940	0.860	0.898	N/A	N/A
7	0.945	0.895	0.938	N/A	N/A
8	0.942	0.883	0.904	N/A	N/A
9	0.942	0.883	0.807	0.997	N/A
10	0.937	0.861	0.901	0.997	N/A
11	0.922	0.839	0.885	N/A	N/A
12	0.912	0.859	0.878	N/A	N/A

Table 5.4.3.3

Marginal Classification Consistency Indices Based on the Domain Scale Scores at the Cut Points: Writ S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.833	0.743	0.989	N/A	N/A
2	0.904	0.769	0.937	N/A	N/A
3	0.933	0.860	0.825	0.998	N/A
4	0.952	0.900	0.622	0.962	0.998
5	0.951	0.907	0.641	0.971	0.995
6	0.912	0.815	0.873	N/A	N/A
7	0.922	0.846	0.889	N/A	N/A
8	0.915	0.833	0.857	N/A	N/A
9	0.914	0.836	0.771	0.995	N/A
10	0.904	0.807	0.848	0.996	N/A
11	0.882	0.787	0.848	N/A	N/A
12	0.874	0.799	0.853	N/A	N/A

5.4.4 Speaking

Table 5.4.4.1

Overall Accuracy and Consistency of Classification Indices: Spek S601 Online

Grade	Accuracy	Consistency
1	0.669	0.573
2	0.649	0.553
3	0.629	0.533
4	0.644	0.534
5	0.624	0.522
6	0.659	0.566
7	0.682	0.575
8	0.631	0.554
9	0.743	0.654
10	0.751	0.656
11	0.725	0.640
12	0.723	0.636

Table 5.4.4.2

Marginal Classification Accuracy Indices Based on the Domain Scale Scores at the Cut Points: Spek S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.918	0.854	0.895	0.996	N/A
2	0.932	0.857	0.859	0.993	N/A
3	0.940	0.875	0.816	0.993	0.998
4	0.958	0.904	0.829	0.949	0.995
5	0.943	0.889	0.814	0.968	0.998
6	0.931	0.876	0.852	0.993	N/A
7	0.928	0.875	0.874	0.996	N/A
8	0.917	0.863	0.844	0.998	N/A
9	0.907	0.880	0.948	N/A	N/A
10	0.911	0.874	0.958	N/A	N/A
11	0.909	0.866	0.943	N/A	N/A
12	0.904	0.841	0.969	N/A	N/A

Table 5.4.4.3

Marginal Classification Consistency Indices Based on the Domain Scale Scores at the Cut Points: Spek S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.881	0.799	0.866	0.995	N/A
2	0.899	0.804	0.824	0.992	N/A
3	0.913	0.817	0.766	0.991	0.998
4	0.937	0.866	0.761	0.932	0.995
5	0.916	0.849	0.750	0.960	0.998
6	0.901	0.830	0.814	0.993	N/A
7	0.897	0.824	0.823	0.995	N/A
8	0.882	0.812	0.819	0.997	N/A
9	0.869	0.828	0.919	N/A	N/A
10	0.873	0.819	0.928	N/A	N/A
11	0.872	0.812	0.922	N/A	N/A
12	0.864	0.787	0.951	N/A	N/A

5.5 Reliabilities of Students' Composite Scale Scores

The reliabilities of the ACCESS composite scale scores indicate the consistency of those scores over replications of the testing procedure. Because the domains that make up the composites consist of different test items, and because items from different domains may measure different abilities (even though items within the domain are assumed to measure a single ability), a traditional internal consistency index such as Cronbach's coefficient alpha is not appropriate, since statisticians who devised such indices assumed that items in a test measure similar ability. It is more appropriate to report a stratified Cronbach's coefficient alpha (Feldt & Brennan, 1989), which measures consistency in students' composite scale scores when those scores are based on students' responses to sets of items that measure different abilities. A stratified alpha is a weighted average of Cronbach's coefficient alphas for item sets that differ in the maximum score points or "strata." Stratified alpha is a reliability estimate computed by dividing the test into components (strata), computing a Cronbach's coefficient alpha separately for the scale scores for each component, and then using the results to estimate a reliability coefficient for the composite scale scores.

In computing the stratified Cronbach's coefficient alphas for ACCESS composite scale scores, we treated each domain that makes up a composite as a separate component (or stratum). For example, when computing the stratified Cronbach's coefficient alphas for students' Literacy scale scores, we entered the variances of the students' scale scores for two components (i.e., Reading and Writing) and the weights of those two components. The stratified Cronbach's coefficient alpha is interpreted like other traditional internal consistency statistics such as Cronbach's coefficient alpha. Like Cronbach's coefficient alpha, a stratified Cronbach's coefficient alpha is an estimate of the proportion of the total variance in the students' composite scale scores that the variance in their true composite scale scores can explain.

Because of the differential weights applied to the ACCESS domains that contribute to the students' composite scale scores, the stratified Cronbach's coefficient alpha is weighted by the contribution that each domain makes to the students' composite scale scores (Kamata, Turhan, & Darandari, 2003; Kane & Case, 2004; Rudner, 2001). Specifically, the formula is

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

where

k = the number of components (domains) j that contribute to the composite

w_j = the weight of component (domain) j

σ_j^2 = the variance of the students' scale scores for component (domain) j

σ_c^2 = the variance of the students' composite scale scores

ρ_j = the reliability coefficient for students' scale scores for component (domain) j.

As is true for the Cronbach's coefficient alpha (see the explanation in Section 5), there is no one set of criteria that the testing community uses when interpreting stratified Cronbach's coefficient alpha values. There is little consensus among the experts in their views of what the acceptable lower limit of the stratified Cronbach's coefficient alpha value should be, or for that matter, how one should interpret various values. This lack of consensus led the authors of the *Standards for Educational and Psychological Measurement* (2014) to conclude, "The choice of [reliability/precision] estimation and the minimum acceptable level for any index remain a matter of professional judgment" (p. 41).

The tables report the stratified Cronbach's coefficient alphas for the students' scale scores for each of the four composites (Oral, Literacy, Comprehension, Overall). The first table for each composite provides stratified Cronbach's coefficient alphas for all students' composite scale scores. The second table for each composite provides the same information for the population of female students and for the population of male students. The third table provides information by ethnicity, for Hispanic and for non-Hispanic students, and the fourth table provides information for the population of students who have an IEP.

The first column of each table shows the grade-level clusters. The tables report the input values that we used to compute the stratified Cronbach's coefficient alphas (i.e., the number of components for each composite, each component's weight, and the variance of the students' scale scores for each component). See Chapter 3 for an explanation of the procedures we used to compute the composite scale scores.

For the students' scale scores in the Listening and Reading domain components, the reliability coefficient is the Rasch student separation reliability coefficient, provided in Section 5.1.

For the students' scale scores in the Writing and Speaking domain components, which have multiple test forms for each grade-level cluster, we derived a single reliability coefficient for the grade-level cluster. To produce this single value, we weighted the Cronbach's coefficient alpha for each of the tiers in the grade-level cluster (provided in Section 5.1) by the number of students who were administered the tier form. We report the weighted average in the tables.

For each relevant domain component, we report the variance of the students' domain scale scores. We also report the variance of the students' composite scale scores. When we computed the variances of the students' domain scale scores and the variances of the students' composite scale scores, we included the students who had valid scores for all four domains.

Finally, the tables present the computed stratified Cronbach's coefficient alphas for students' scale scores for each composite, by grade-level cluster.

Additionally, we used the stratified Cronbach's coefficient alphas, presented in the tables in this section, to produce the **Accuracy and Consistency** classification tables for the composites (Section 5.7). The stratified Cronbach's coefficient alphas for the Oral scale scores computed for all students ranged from 0.90 to 0.91. The stratified Cronbach's coefficient alphas for the Oral scale scores ranged from 0.91 to 0.92 for male students; 0.90 to 0.91 for female students; 0.90 to 0.91 for Hispanic students; 0.89 to 0.91 for non-Hispanic students; and 0.88 to 0.91 for students with an IEP.

The stratified Cronbach's coefficient alphas for the Literacy scale scores computed for all students ranged from 0.87 to 0.89. The stratified Cronbach's coefficient alphas for the Literacy scale scores ranged from 0.87 to 0.89 for male students; 0.85 to 0.89 for female students; 0.87 to 0.89 for Hispanic students; 0.87 to 0.89 for non-Hispanic students; and 0.85 to 0.89 for students with an IEP.

The stratified Cronbach's coefficient alphas for the Comprehension scale scores computed for all students ranged from 0.90 to 0.93. The stratified Cronbach's coefficient alphas for the Comprehension scale scores ranged from 0.90 to 0.93 for male students; 0.89 to 0.93 for female

students; 0.88 to 0.93 for Hispanic students; 0.91 to 0.93 for non-Hispanic students; and 0.87 to 0.91 for students with an IEP.

Since all WIDA states use students' Overall scale scores in making accountability decisions, it is critical that the students' Overall scale score have high reliability. The stratified Cronbach's coefficient alphas for the Overall scale scores computed for all students ranged from 0.93 to 0.94. The stratified Cronbach's coefficient alphas for the Overall scale scores ranged from 0.93 to 0.94 for male students; 0.92 to 0.93 for female students; 0.92 to 0.94 for Hispanic students; 0.92 to 0.93 for non-Hispanic students; and 0.91 to 0.93 for students with an IEP.

5.5.1 Oral

Table 5.5.1.1

Reliabilities of Composite Scale Scores: Oral S601 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.50	3057.74	0.88
	Speaking	0.50	3316.05	0.84
	Oral		2542.75	0.91
2-3	Listening	0.50	3528.70	0.88
	Speaking	0.50	3600.55	0.84
	Oral		2895.12	0.91
4-5	Listening	0.50	2942.32	0.85
	Speaking	0.50	3543.14	0.84
	Oral		2639.01	0.90
6-8	Listening	0.50	2350.90	0.85
	Speaking	0.50	3397.22	0.85
	Oral		2293.95	0.91
9-12	Listening	0.50	2388.21	0.86
	Speaking	0.50	3868.53	0.86
	Oral		2488.40	0.91

Table 5.5.1.2

Reliabilities of Composite Scale Scores: Oral S601 Online by Gender

			Female		Male	
Cluster	Component	Weight	Variance	Reliability	Variance	Reliability
1	Listening	0.50	2974.67	0.87	3120.53	0.88
	Speaking	0.50	3340.57	0.84	3277.61	0.84
	Oral		2523.57	0.91	2548.25	0.91
2-3	Listening	0.50	3315.17	0.87	3705.44	0.89
	Speaking	0.50	3592.20	0.84	3609.25	0.84
	Oral		2801.91	0.91	2978.31	0.92
4-5	Listening	0.50	2795.74	0.85	3062.41	0.86
	Speaking	0.50	3550.20	0.84	3557.81	0.84
	Oral		2569.09	0.90	2710.07	0.91
6-8	Listening	0.50	2293.01	0.85	2397.49	0.85
	Speaking	0.50	3480.70	0.85	3315.38	0.85
	Oral		2297.10	0.91	2288.73	0.91
9-12	Listening	0.50	2307.78	0.85	2443.51	0.86
	Speaking	0.50	3858.66	0.86	3885.07	0.87
	Oral		2457.49	0.91	2510.12	0.91

Table 5.5.1.3

Reliabilities of Composite Scale Scores: Oral S601 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.50	2971.96	0.88	3032.08	0.87
	Speaking	0.50	3277.28	0.84	3125.57	0.83
	Oral		2479.66	0.91	2436.06	0.90
2-3	Listening	0.50	3346.22	0.88	3604.80	0.88
	Speaking	0.50	3656.24	0.84	3232.81	0.83
	Oral		2829.63	0.91	2757.15	0.91
4-5	Listening	0.50	2818.94	0.85	2814.36	0.83
	Speaking	0.50	3524.38	0.84	3041.39	0.82
	Oral		2557.54	0.90	2353.92	0.89
6-8	Listening	0.50	2282.68	0.85	2257.30	0.84
	Speaking	0.50	3362.98	0.85	2954.00	0.83
	Oral		2231.54	0.91	2068.95	0.90
9-12	Listening	0.50	2367.01	0.85	2084.97	0.83
	Speaking	0.50	3889.18	0.87	3270.57	0.84
	Oral		2472.31	0.91	2076.21	0.89

Table 5.5.1.4

Reliabilities of Composite Scale Scores: Oral S601 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.50	3114.81	0.88
	Speaking	0.50	3431.98	0.84
	Oral		2577.58	0.91
2-3	Listening	0.50	3240.61	0.88
	Speaking	0.50	3494.11	0.83
	Oral		2642.82	0.91
4-5	Listening	0.50	2376.72	0.84
	Speaking	0.50	2986.60	0.83
	Oral		2027.20	0.89
6-8	Listening	0.50	1751.36	0.81
	Speaking	0.50	2796.42	0.84
	Oral		1666.06	0.88
9-12	Listening	0.50	1532.30	0.79
	Speaking	0.50	3439.46	0.87
	Oral		1766.27	0.89

5.5.2 Literacy

Table 5.5.2.1

Reliabilities of Composite Scale Scores: Litr S601 Online

Cluster	Component	Weight	Variance	Reliability
1	Reading	0.50	780.31	0.84
	Writing	0.50	2405.67	0.85
	Literacy		1122.65	0.89
2-3	Reading	0.50	967.74	0.87
	Writing	0.50	2318.44	0.81
	Literacy		1238.04	0.88
4-5	Reading	0.50	1103.07	0.89
	Writing	0.50	2664.87	0.75
	Literacy		1513.06	0.87
6-8	Reading	0.50	1168.67	0.89
	Writing	0.50	1685.82	0.77
	Literacy		1167.84	0.89
9-12	Reading	0.50	1319.07	0.90
	Writing	0.50	1590.32	0.74
	Literacy		1152.09	0.88

Table 5.5.2.2

Reliabilities of Composite Scale Scores: Litr S601 Online by Gender

			Female		Male	
Cluster	Component	Weight	Variance	Reliability	Variance	Reliability
1	Reading	0.50	792.13	0.84	771.43	0.84
	Writing	0.50	2309.55	0.84	2510.68	0.85
	Literacy		1109.62	0.89	1144.74	0.89
2-3	Reading	0.50	936.77	0.87	992.70	0.87
	Writing	0.50	2272.24	0.80	2334.87	0.82
	Literacy		1217.62	0.88	1247.90	0.89
4-5	Reading	0.50	1037.34	0.88	1147.93	0.89
	Writing	0.50	2564.05	0.72	2750.12	0.76
	Literacy		1453.83	0.85	1561.30	0.87
6-8	Reading	0.50	1145.43	0.89	1183.91	0.89
	Writing	0.50	1630.98	0.75	1729.24	0.79
	Literacy		1140.32	0.88	1188.89	0.89
9-12	Reading	0.50	1246.00	0.90	1356.96	0.90
	Writing	0.50	1558.66	0.73	1625.65	0.76
	Literacy		1112.72	0.88	1177.62	0.89

Table 5.5.2.3

Reliabilities of Composite Scale Scores: Litr S601 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Reading	0.50	621.42	0.80	1017.83	0.87
	Writing	0.50	2315.40	0.85	2167.77	0.82
	Literacy		970.96	0.88	1215.14	0.89
2-3	Reading	0.50	873.50	0.86	1078.78	0.88
	Writing	0.50	2354.53	0.82	1948.92	0.77
	Literacy		1186.47	0.88	1177.12	0.88
4-5	Reading	0.50	1049.46	0.88	1137.49	0.89
	Writing	0.50	2628.48	0.75	2272.34	0.73
	Literacy		1467.12	0.87	1388.31	0.87
6-8	Reading	0.50	1113.42	0.88	1228.44	0.89
	Writing	0.50	1655.15	0.77	1472.33	0.75
	Literacy		1124.05	0.89	1117.15	0.89
9-12	Reading	0.50	1242.25	0.90	1398.22	0.91
	Writing	0.50	1565.03	0.75	1413.53	0.71
	Literacy		1106.40	0.88	1093.95	0.88

Table 5.5.2.4

Reliabilities of Composite Scale Scores: Litr S601 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Reading	0.50	595.19	0.79
	Writing	0.50	2823.68	0.87
	Literacy		1097.98	0.89
2-3	Reading	0.50	739.84	0.83
	Writing	0.50	2339.05	0.85
	Literacy		1050.27	0.89
4-5	Reading	0.50	981.86	0.87
	Writing	0.50	2391.93	0.80
	Literacy		1290.17	0.88
6-8	Reading	0.50	885.30	0.85
	Writing	0.50	1268.28	0.79
	Literacy		820.39	0.88
9-12	Reading	0.50	972.47	0.87
	Writing	0.50	1146.97	0.73
	Literacy		733.70	0.85

5.5.3 Comprehension

Table 5.5.3.1

Reliabilities of Composite Scale Scores: Cphn S601 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.30	3057.74	0.88
	Reading	0.70	780.31	0.84
	Comprehension		907.28	0.90
2-3	Listening	0.30	3528.70	0.88
	Reading	0.70	967.74	0.87
	Comprehension		1243.92	0.92
4-5	Listening	0.30	2942.32	0.85
	Reading	0.70	1103.07	0.89
	Comprehension		1311.18	0.92
6-8	Listening	0.30	2350.90	0.85
	Reading	0.70	1168.67	0.89
	Comprehension		1251.81	0.92
9-12	Listening	0.30	2388.21	0.86
	Reading	0.70	1319.07	0.90
	Comprehension		1394.90	0.93

Table 5.5.3.2

Reliabilities of Composite Scale Scores: Cphn S601 Online by Gender

			Female		Male	
Cluster	Component	Weight	Variance	Reliability	Variance	Reliability
1	Listening	0.30	2974.67	0.87	3120.53	0.88
	Reading	0.70	792.13	0.84	771.43	0.84
	Comprehension		904.45	0.89	910.33	0.90
2-3	Listening	0.30	3315.17	0.87	3705.44	0.89
	Reading	0.70	936.77	0.87	992.70	0.87
	Comprehension		1189.20	0.92	1287.66	0.92
4-5	Listening	0.30	2795.74	0.85	3062.41	0.86
	Reading	0.70	1037.34	0.88	1147.93	0.89
	Comprehension		1237.23	0.92	1364.21	0.93
6-8	Listening	0.30	2293.01	0.85	2397.49	0.85
	Reading	0.70	1145.43	0.89	1183.91	0.89
	Comprehension		1231.47	0.92	1268.21	0.92
9-12	Listening	0.30	2307.78	0.85	2443.51	0.86
	Reading	0.70	1246.00	0.90	1356.96	0.90
	Comprehension		1335.18	0.93	1428.30	0.93

Table 5.5.3.3

Reliabilities of Composite Scale Scores: Cphn S601 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.30	2971.96	0.88	3032.08	0.87
	Reading	0.70	621.42	0.80	1017.83	0.87
	Comprehension		752.49	0.88	1115.67	0.91
2-3	Listening	0.30	3346.22	0.88	3604.80	0.88
	Reading	0.70	873.50	0.86	1078.78	0.88
	Comprehension		1116.30	0.91	1378.69	0.93
4-5	Listening	0.30	2818.94	0.85	2814.36	0.83
	Reading	0.70	1049.46	0.88	1137.49	0.89
	Comprehension		1237.91	0.92	1327.42	0.92
6-8	Listening	0.30	2282.68	0.85	2257.30	0.84
	Reading	0.70	1113.42	0.88	1228.44	0.89
	Comprehension		1191.00	0.92	1283.73	0.92
9-12	Listening	0.30	2367.01	0.85	2084.97	0.83
	Reading	0.70	1242.25	0.90	1398.22	0.91
	Comprehension		1330.17	0.93	1383.07	0.93

Table 5.5.3.4

Reliabilities of Composite Scale Scores: Cphn S601 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.30	3114.81	0.88
	Reading	0.70	595.19	0.79
	Comprehension		737.11	0.87
2-3	Listening	0.30	3240.61	0.88
	Reading	0.70	739.84	0.83
	Comprehension		935.86	0.90
4-5	Listening	0.30	2376.72	0.84
	Reading	0.70	981.86	0.87
	Comprehension		1048.17	0.91
6-8	Listening	0.30	1751.36	0.81
	Reading	0.70	885.30	0.85
	Comprehension		877.96	0.89
9-12	Listening	0.30	1532.30	0.79
	Reading	0.70	972.47	0.87
	Comprehension		905.33	0.90

5.5.4 Overall

Table 5.5.4.1

Reliabilities of Composite Scale Scores: Over S601 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.15	3057.74	0.88
	Reading	0.35	780.31	0.84
	Writing	0.35	2405.67	0.85
	Speaking	0.15	3316.05	0.84
	Overall Composite		1189.52	0.93
2-3	Listening	0.15	3528.70	0.88
	Reading	0.35	967.74	0.87
	Writing	0.35	2318.44	0.81
	Speaking	0.15	3600.55	0.84
	Overall Composite		1426.88	0.94
4-5	Listening	0.15	2942.32	0.85
	Reading	0.35	1103.07	0.89
	Writing	0.35	2664.87	0.75
	Speaking	0.15	3543.14	0.84
	Overall Composite		1613.06	0.93
6-8	Listening	0.15	2350.90	0.85
	Reading	0.35	1168.67	0.89
	Writing	0.35	1685.82	0.77
	Speaking	0.15	3397.22	0.85
	Overall Composite		1300.92	0.94
9-12	Listening	0.15	2388.21	0.86
	Reading	0.35	1319.07	0.90
	Writing	0.35	1590.32	0.74
	Speaking	0.15	3868.53	0.86
	Overall Composite		1349.14	0.94

Table 5.5.4.2

Reliabilities of Composite Scale Scores: Over S601 Online by Gender

Cluster	Component	Weight	Female		Male	
			Variance	Reliability	Variance	Reliability
1	Listening	0.15	2974.67	0.87	3120.53	0.88
	Reading	0.35	792.13	0.84	771.43	0.84
	Writing	0.35	2309.55	0.84	2510.68	0.85
	Speaking	0.15	3340.57	0.84	3277.61	0.84
	Overall Composite		1175.74	0.93	1207.04	0.93
2-3	Listening	0.15	3315.17	0.87	3705.44	0.89
	Reading	0.35	936.77	0.87	992.70	0.87
	Writing	0.35	2272.24	0.80	2334.87	0.82
	Speaking	0.15	3592.20	0.84	3609.25	0.84
	Overall Composite		1396.00	0.93	1447.30	0.94
4-5	Listening	0.15	2795.74	0.85	3062.41	0.86
	Reading	0.35	1037.34	0.88	1147.93	0.89
	Writing	0.35	2564.05	0.72	2750.12	0.76
	Speaking	0.15	3550.20	0.84	3557.81	0.84
	Overall Composite		1555.85	0.92	1663.34	0.93
6-8	Listening	0.15	2293.01	0.85	2397.49	0.85
	Reading	0.35	1145.43	0.89	1183.91	0.89
	Writing	0.35	1630.98	0.75	1729.24	0.79
	Speaking	0.15	3480.70	0.85	3315.38	0.85
	Overall Composite		1285.26	0.93	1314.53	0.94
9-12	Listening	0.15	2307.78	0.85	2443.51	0.86
	Reading	0.35	1246.00	0.90	1356.96	0.90
	Writing	0.35	1558.66	0.73	1625.65	0.76
	Speaking	0.15	3858.66	0.86	3885.07	0.87
	Overall Composite		1316.71	0.93	1369.84	0.94

Table 5.5.4.3

Reliabilities of Composite Scale Scores: Over S601 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.15	2971.96	0.88	3032.08	0.87
	Reading	0.35	621.42	0.80	1017.83	0.87
	Writing	0.35	2315.40	0.85	2167.77	0.82
	Speaking	0.15	3277.28	0.84	3125.57	0.83
	Overall Composite		1053.74	0.93	1253.24	0.93
2-3	Listening	0.15	3346.22	0.88	3604.80	0.88
	Reading	0.35	873.50	0.86	1078.78	0.88
	Writing	0.35	2354.53	0.82	1948.92	0.77
	Speaking	0.15	3656.24	0.84	3232.81	0.83
	Overall Composite		1363.25	0.93	1366.17	0.93
4-5	Listening	0.15	2818.94	0.85	2814.36	0.83
	Reading	0.35	1049.46	0.88	1137.49	0.89
	Writing	0.35	2628.48	0.75	2272.34	0.73
	Speaking	0.15	3524.38	0.84	3041.39	0.82
	Overall Composite		1554.39	0.92	1458.24	0.92
6-8	Listening	0.15	2282.68	0.85	2257.30	0.84
	Reading	0.35	1113.42	0.88	1228.44	0.89
	Writing	0.35	1655.15	0.77	1472.33	0.75
	Speaking	0.15	3362.98	0.85	2954.00	0.83
	Overall Composite		1248.48	0.94	1214.46	0.93
9-12	Listening	0.15	2367.01	0.85	2084.97	0.83
	Reading	0.35	1242.25	0.90	1398.22	0.91
	Writing	0.35	1565.03	0.75	1413.53	0.71
	Speaking	0.15	3889.18	0.87	3270.57	0.84
	Overall Composite		1308.96	0.94	1205.95	0.93

Table 5.5.4.4

Reliabilities of Composite Scale Scores: Over S601 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.15	3114.81	0.88
	Reading	0.35	595.19	0.79
	Writing	0.35	2823.68	0.87
	Speaking	0.15	3431.98	0.84
	Overall Composite		1121.22	0.93
2-3	Listening	0.15	3240.61	0.88
	Reading	0.35	739.84	0.83
	Writing	0.35	2339.05	0.85
	Speaking	0.15	3494.11	0.83
	Overall Composite		1165.18	0.93
4-5	Listening	0.15	2376.72	0.84
	Reading	0.35	981.86	0.87
	Writing	0.35	2391.93	0.80
	Speaking	0.15	2986.60	0.83
	Overall Composite		1243.39	0.92
6-8	Listening	0.15	1751.36	0.81
	Reading	0.35	885.30	0.85
	Writing	0.35	1268.28	0.79
	Speaking	0.15	2796.42	0.84
	Overall Composite		855.07	0.92
9-12	Listening	0.15	1532.30	0.79
	Reading	0.35	972.47	0.87
	Writing	0.35	1146.97	0.73
	Speaking	0.15	3439.46	0.87
	Overall Composite		820.82	0.91

5.6 CSEMs for the Students' Composite Scale Scores

CSEMs for the four ACCESS composite scale scores provide test users with a benchmark indicating how free a student's composite scale score is from measurement errors at different WIDA proficiency levels. Due to the differential weights applied to different ACCESS domains (see the introduction to Section 3 for weighting conventions), WIDA estimates the CSEMs using a procedure that is based on IRT (Lord, 1980) and developed by Price, Lurie, Raju, Wilkins, and Zhu (2006). Price et al. (2006) extended the work by Lord (1980) and Kolen, Hanson, and Brennan (1992) in estimating the CSEMs of students' composite scale scores consisting of components. The basic premise of this procedure is that one can estimate empirically the CSEM for a student's weighted composite scale score using the IRT-based CSEMs for each student's component scale scores and the weights associated with the components. We used this method to estimate the CSEMs for ACCESS composite scale scores by treating the ACCESS domains as components.

We used a three-step process to derive the CSEM for each ACCESS composite scale score. We calculated a unique CSEM for each composite scale score by grade. Since this procedure relies on empirical student data, which are subject to year-to-year fluctuations, we used all population student data from all previous three ACCESS 2.0 series in our calculations to obtain more stable estimates than using data from just a single series.

Step 1. Since we calibrated ACCESS domains separately, measurement errors associated with each of the ACCESS domains, as expressed in the CSEM, were independent of each other. Therefore, we estimated the CSEM for a student's composite scale score x, SEM_x, using the equation derived by Price et al. (2006):

$$SEM_x = \sqrt{W_1^2 SEM_1^2 + W_2^2 SEM_2^2 + W_3^2 SEM_3^2 + \cdots + W_k^2 SEM_k^2}$$

Where SEM_i^2 is the student's IRT-based score error variance or the squared CSEM for the student's scale score for ACCESS domain *i*, and W_i is the weight applied to domain *i*, for *i*=1,…,*k*.

Step 2. Due to the differential weights applied to different ACCESS domains, two students whose weighted domain scale scores are the same may have composite scale scores with

different CSEMs; therefore, we instituted an additional step to obtain a unique CSEM value for each composite scale score. Specifically, we estimated the expected value of the CSEM functions for a composite scale score using a regression approach, and we reported this expected value as the CSEM for that composite scale score.

Step 3. We applied a linear smoothing procedure to derive the CSEMs for composite scale scores that we did not observe in the data.

The figures in this section show graphically the CSEMs for various composite scale scores by grade level. The students' composite scale scores appear on the horizontal axis, and the corresponding CSEMs appear on the vertical axis. Each point in a figure represents a student in the dataset, showing the relationship between the CSEM and that student's composite scale score. We did not plot values for students who received the lowest possible scale scores for any ACCESS domains, as it is not possible to compute accurately the CSEM for these students' scale scores. For grade-level clusters with multiple grades, we use different colors in the figures to represent students in different grades.

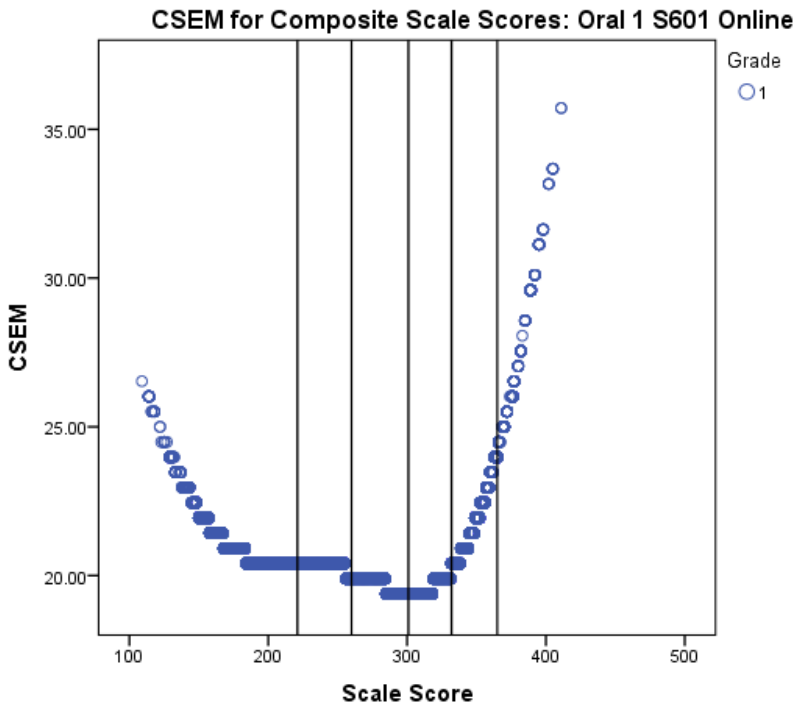
The five vertical lines in the figure indicate the five ACCESS composite scale score cut points for the highest grade in the grade-level cluster for the test form, dividing the figure into six sections representing the six WIDA proficiency levels.

Smaller CSEM values indicate less measurement error (i.e., greater accuracy in measurement). In general, these figures show that the CSEMs are smaller and fairly constant in the middle of the composite scale score range but larger and more variable for extreme low and high composite scale scores. This is to be expected, since we used an IRT approach when scaling ACCESS, which typically produces larger CSEMs for scale scores that are at the lower and the higher ends of the scale score range. In addition, because students exit the EL program when they demonstrate that they are English language proficient, the number of students whose composite scale scores are at the extreme high end of the score range is typically small, as compared to the number of students whose composite scale scores are in the middle of the score range. Therefore, the measurement errors associated with the composite scale scores at the extreme high end of the score range tend to be larger since the calculation of these scale scores is based on the test performances of fewer students.

5.6.1 Oral

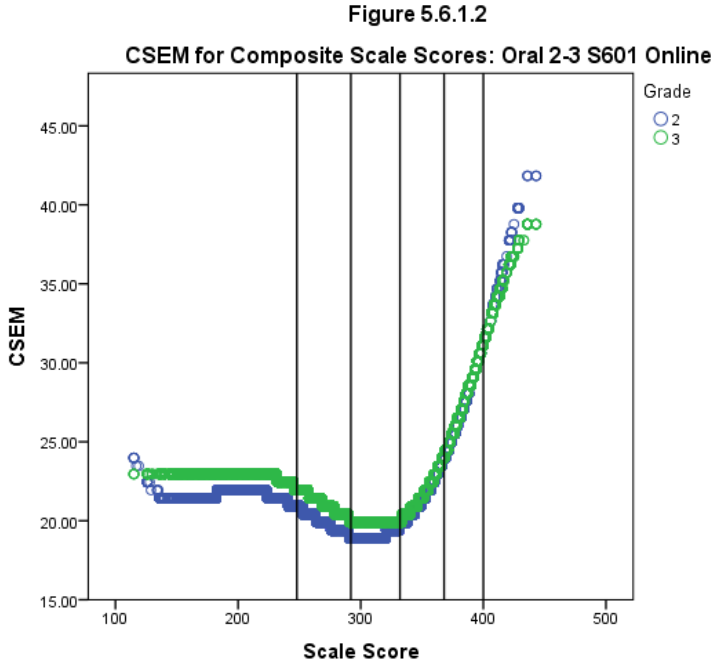
5.6.1.1 Grade 1

Figure 5.6.1.1



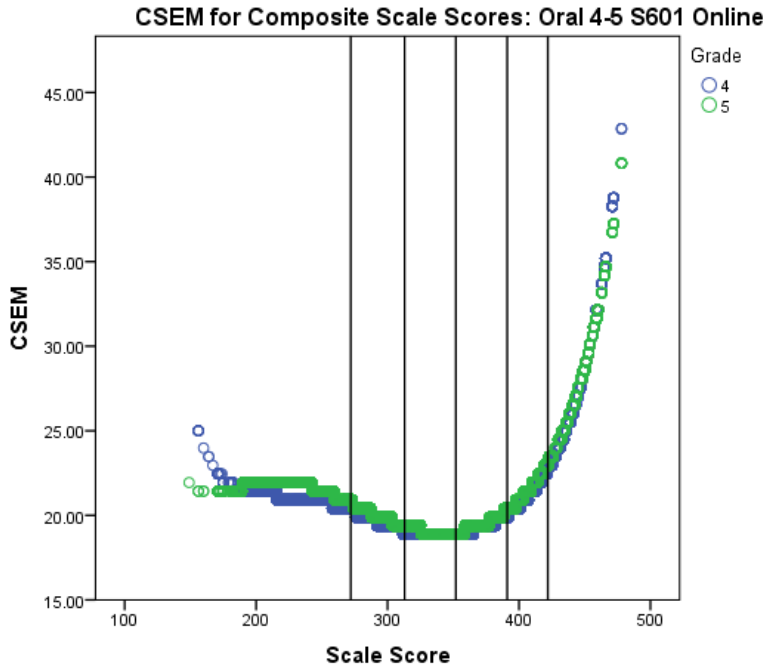
5.6.1.2 Grades 2-3

Figure 5.6.1.2



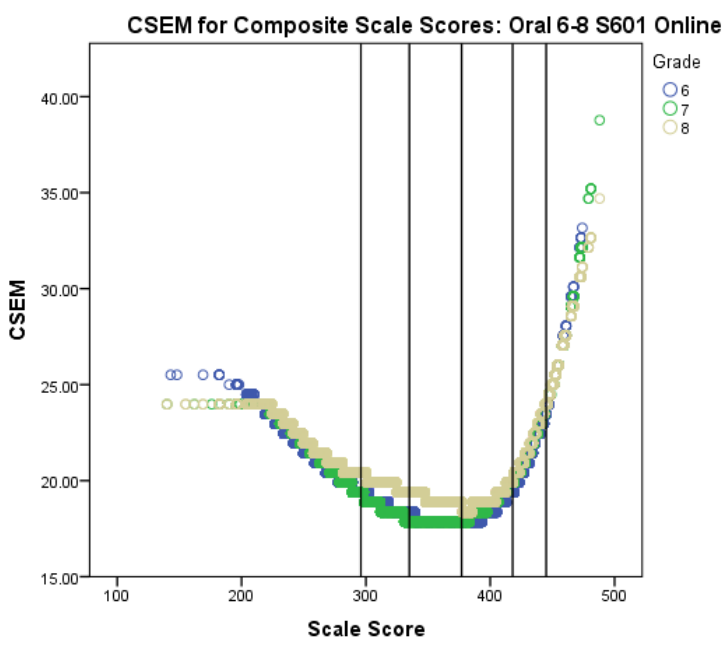
5.6.1.3 Grades 4-5

Figure 5.6.1.3



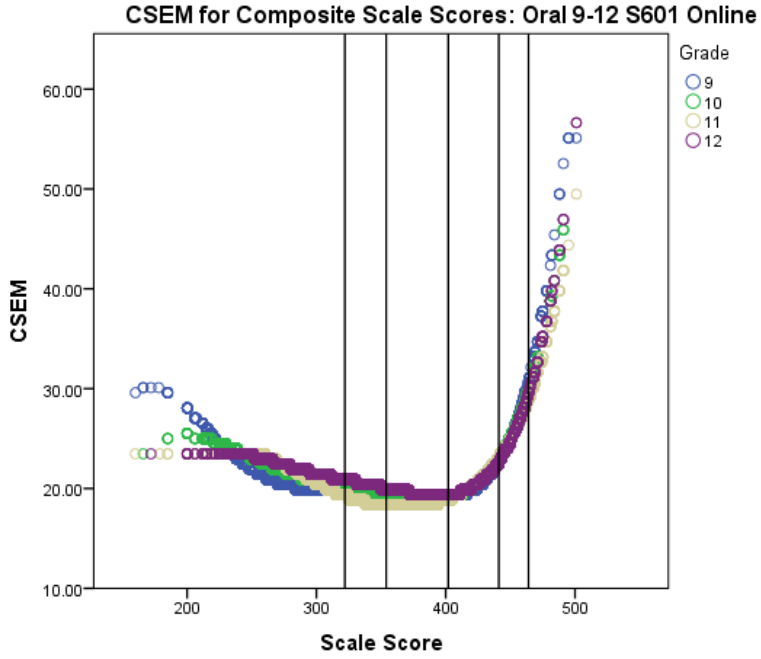
5.6.1.4 Grades 6-8

Figure 5.6.1.4



5.6.1.5 Grades 9-12

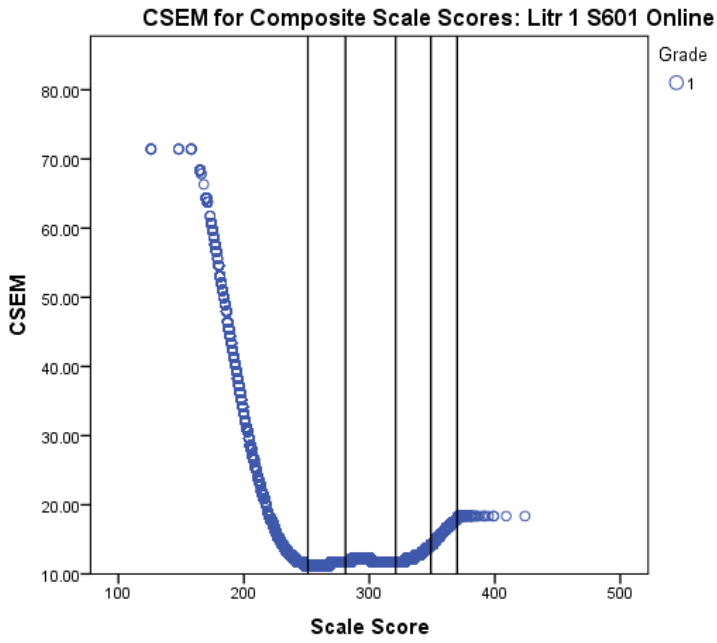
Figure 5.6.1.5



5.6.2 Literacy

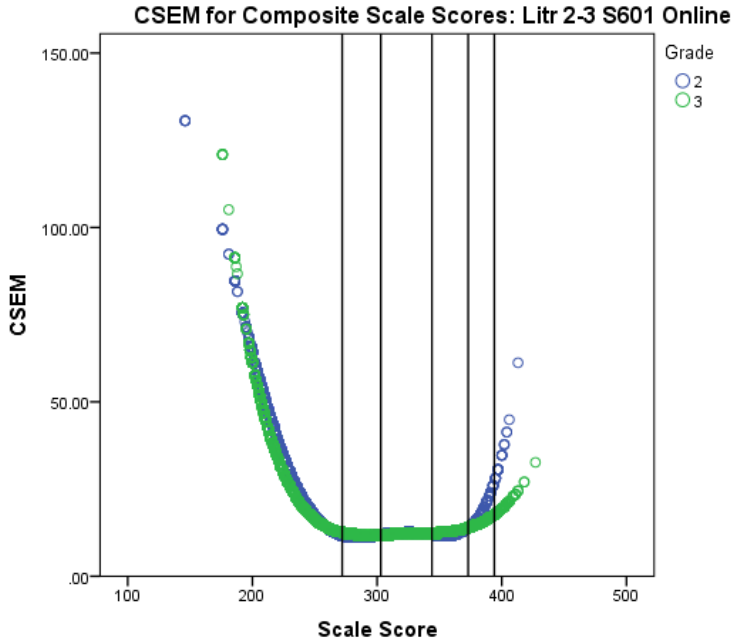
5.6.2.1 Grade 1

Figure 5.6.2.1



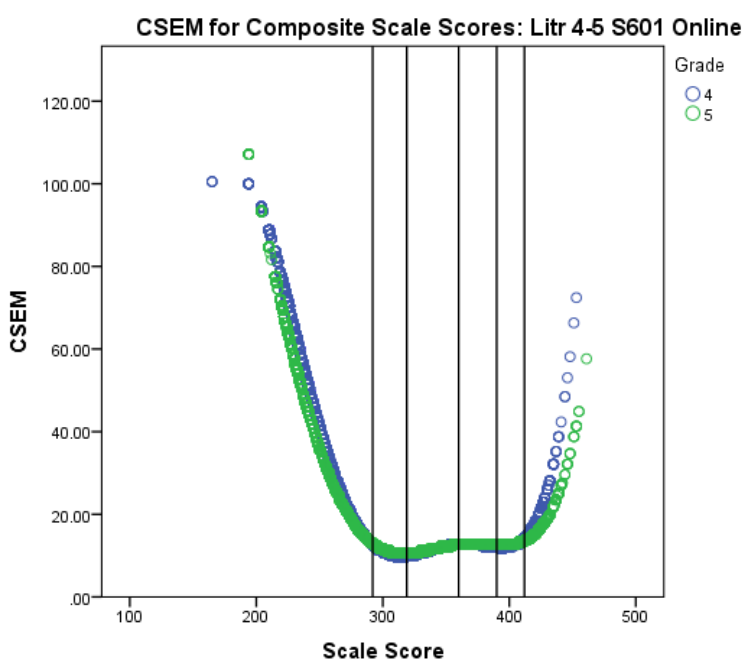
5.6.2.2 Grades 2-3

Figure 5.6.2.2



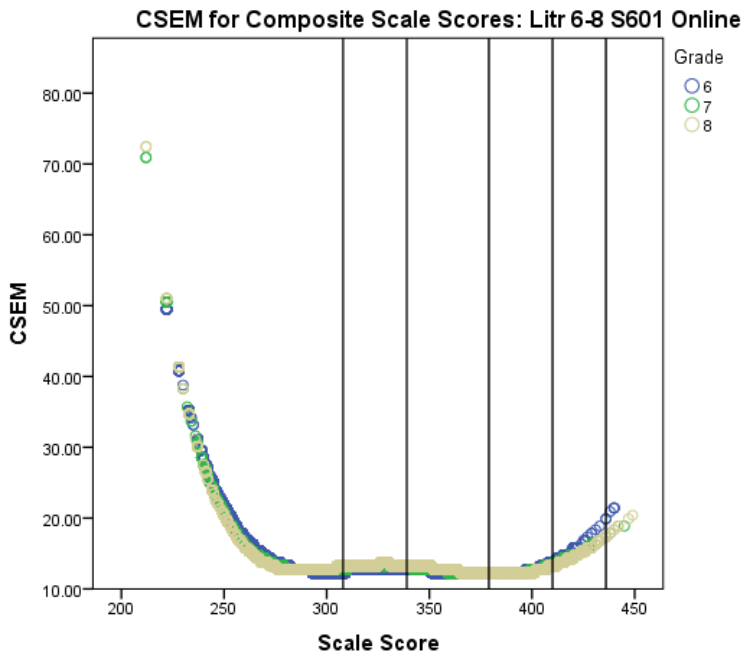
5.6.2.3 Grades 4-5

Figure 5.6.2.3



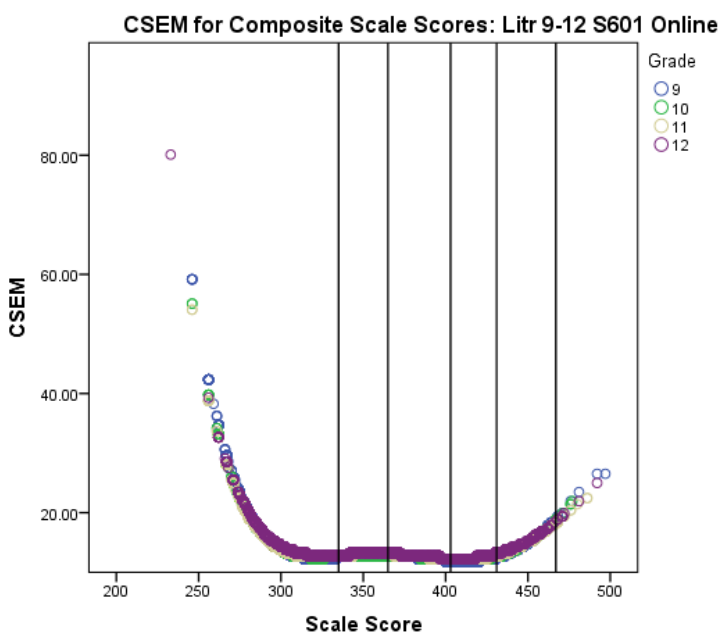
5.6.2.4 Grades 6-8

Figure 5.6.2.4



5.6.2.5 Grades 9-12

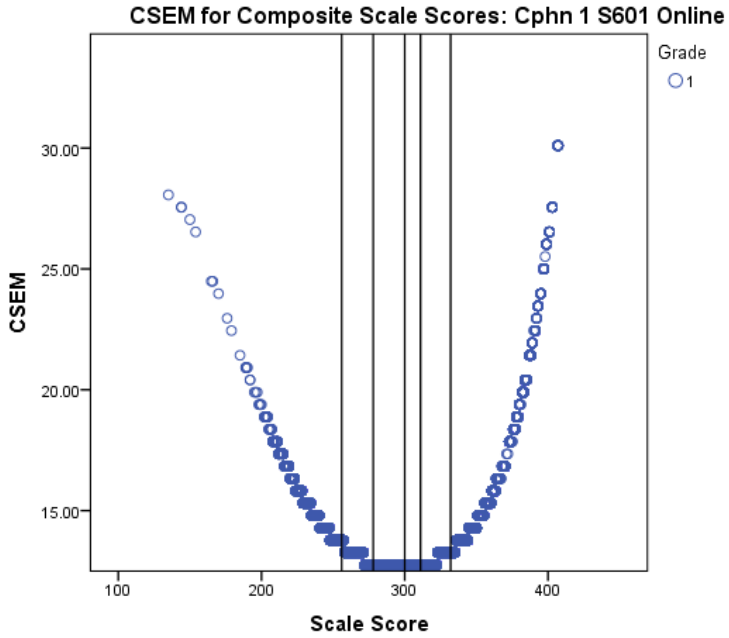
Figure 5.6.2.5



5.6.3 Comprehension

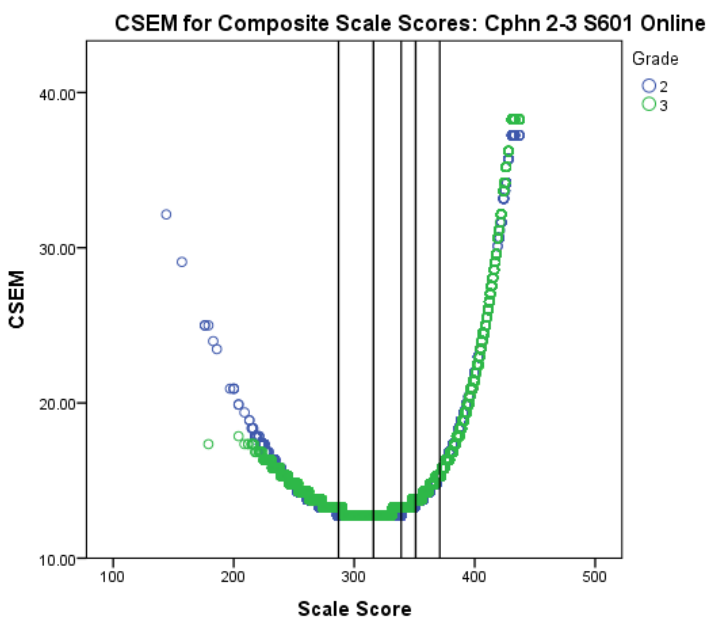
5.6.3.1 Grade 1

Figure 5.6.3.1



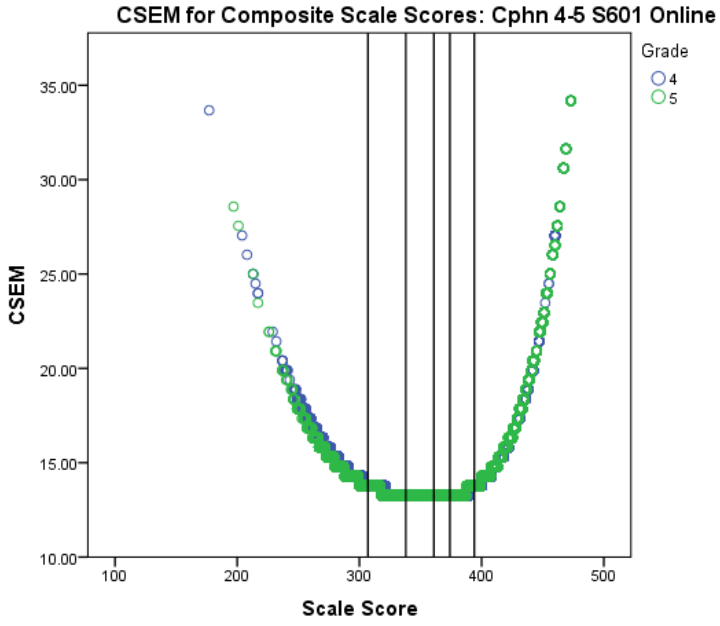
5.6.3.2 Grades 2-3

Figure 5.6.3.2



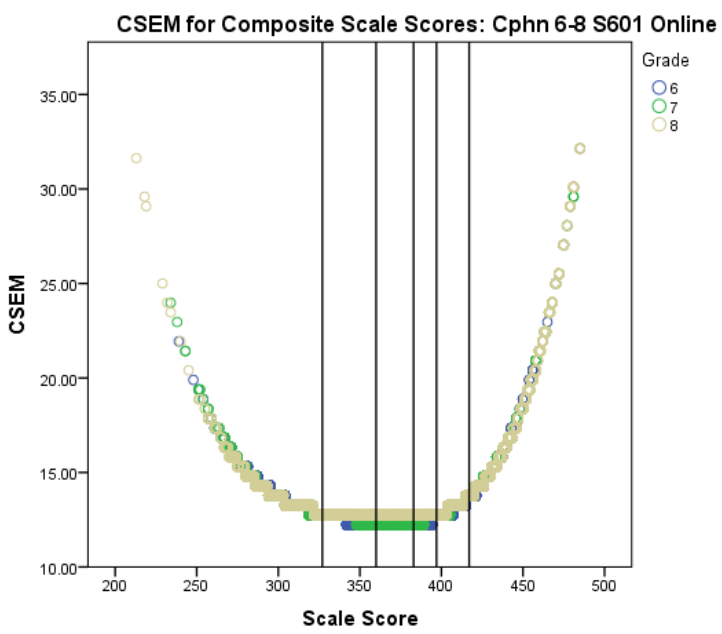
5.6.3.3 Grades 4-5

Figure 5.6.3.3



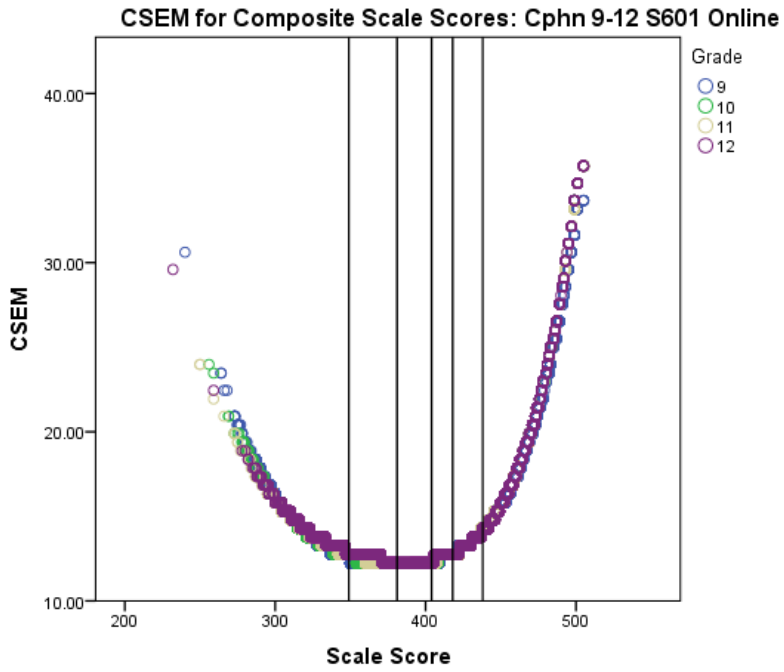
5.6.3.4 Grades 6-8

Figure 5.6.3.4



5.6.3.5 Grades 9-12

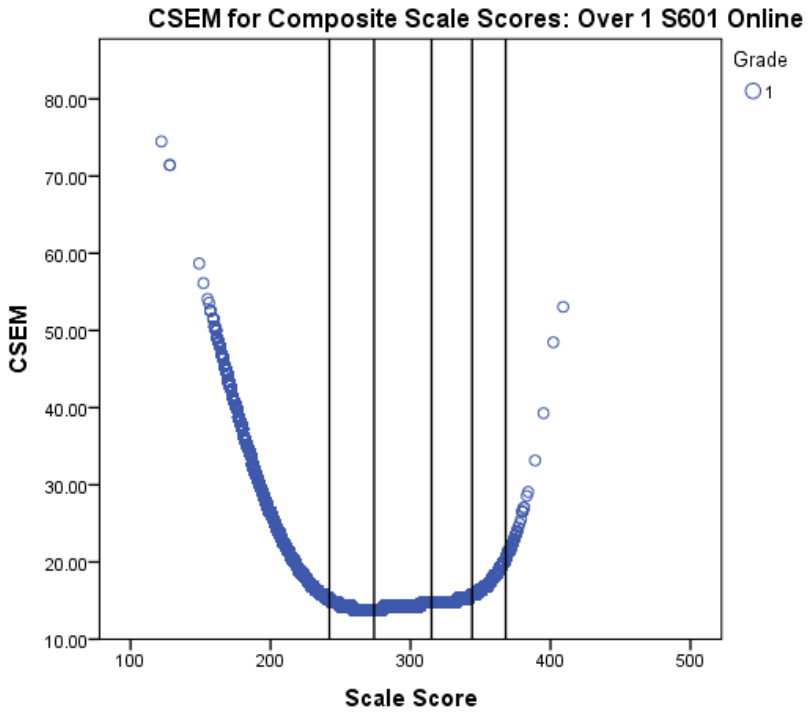
Figure 5.6.3.5



5.6.4 Overall

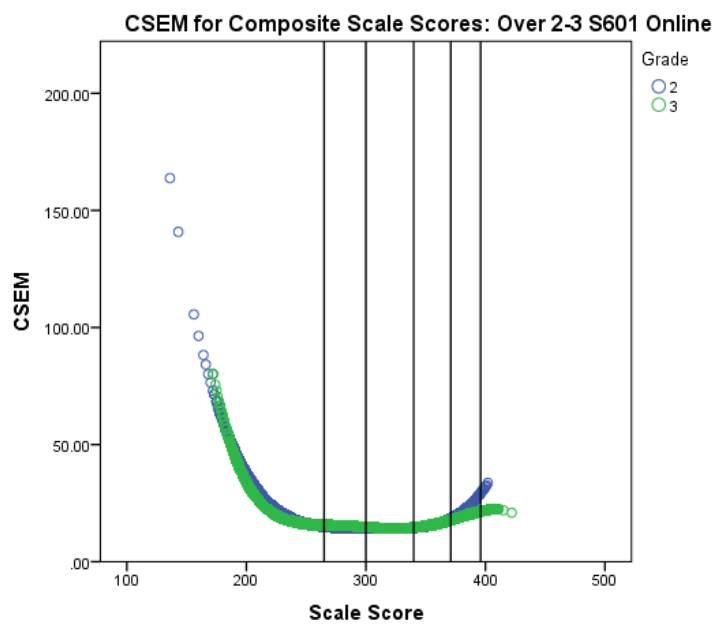
5.6.4.1 Grade 1

Figure 5.6.4.1



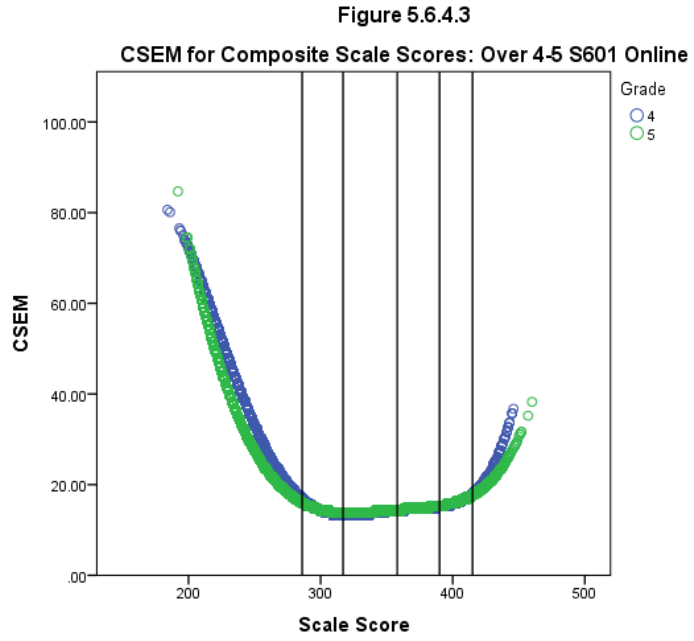
5.6.4.2 Grades 2-3

Figure 5.6.4.2



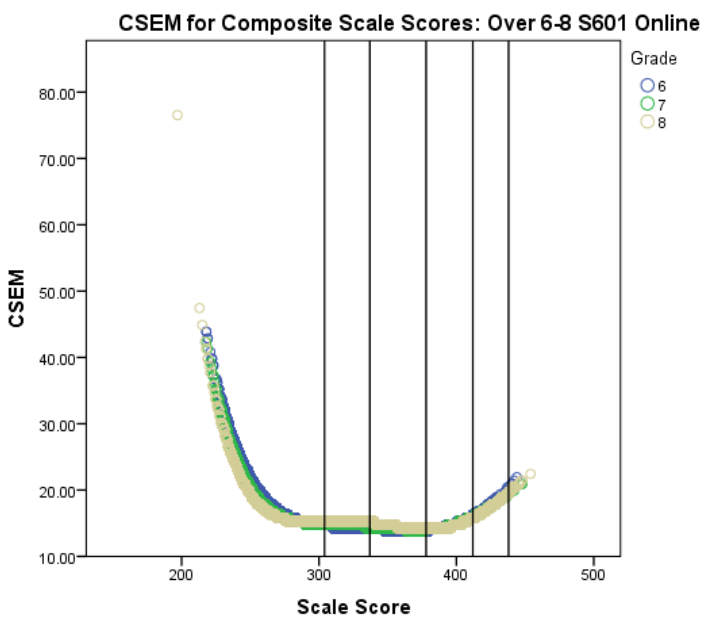
5.6.4.3 Grades 4-5

Figure 5.6.4.3



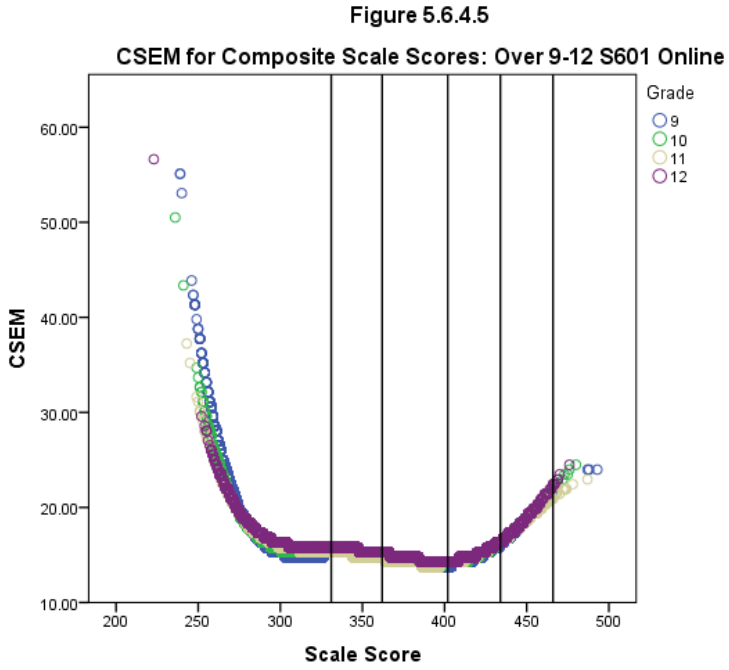
5.6.4.4 Grades 6-8

Figure 5.6.4.4



5.6.4.5 Grades 9-12

Figure 5.6.4.5



5.7 Accuracy and Consistency of Composites

One of the main purposes of the WIDA ACCESS program is to identify the English language proficiency level of students with respect to the WIDA ELD Standards. Because of the emphasis on the classification of student performance, a question of interest is how accurately and consistently the ACCESS composite scale scores can classify students into WIDA proficiency categories determined by the 2016 ACCESS standard-setting process (Cook & MacGregor, 2017). Although states in the WIDA Consortium take into consideration one or more of the domain and composite scale scores when making accountability decisions, all WIDA Consortium states use the **Overall composite scale score** as the primary score when making classification decisions about students. Therefore, it is especially important to examine the accuracy and consistency of the classifications based on the Overall composite scale scores to help test users and policy makers judge the utility of this information and to make decisions about score reporting (American Educational Research Association et al., 2014). The analyses utilize the methods that Livingston and Lewis (1995) and Young and Yoon (1998) outlined, as implemented in the software program BB-CLASS (Brennan, 2004; cf. also Lee et al., 2002).

The method and descriptions of the classification accuracy and consistency indices reported in this section appear in detail in Section 5.4. The only substantive methodological difference between the estimation of the classification accuracy and consistency of the domain scale scores versus the composite scale scores is that to estimate the classification accuracy and consistency of the composite scale scores, we first estimate the reliability of the composite scale scores using a stratified Cronbach's coefficient alpha, as described in Section 5.4.

For each composite, we present three tables. The first reports the overall accuracy and the overall consistency indices for each grade. The second reports the marginal classification accuracy indices based on the composite scale scores at the cut points for each grade. The third reports the marginal classification consistency indices based on the composite scale scores at the cut points for each grade.

If we could not estimate the overall and marginal classification accuracy and consistency indices because there were fewer than 200 students in the proficiency level, we collapsed the affected proficiency level with the level below it and placed 'N/A' in the table for the affected proficiency level.

As noted in Section 5.4, assessment experts have issued very little guidance to aid in making judgments about the ideal or expected levels of decision consistency and accuracy needed for educational assessments. To help test users and policy makers interpret the results from our analyses, we report for each composite the range of these indices, highlighting the grade with the lowest classification accuracy and consistency indices for that composite. Since overall accuracy and consistency indices are summaries of the degree of classification accuracy and consistency for the composite scale scores across all proficiency level cut points, we also examine the marginal classification accuracy and consistency indices for these grades to identify the specific source(s) of low classification accuracy and consistency.

For the Oral composite, as shown in Table 5.7.1.1, the overall classification accuracy indices ranged from 0.644 to 0.755, and the overall classification consistency indices ranged from 0.532 to 0.664 across grades. The lowest overall classification accuracy and consistency indices were found for students in Grade 5.

For the Literacy composite, as shown in Table 5.7.2.1, the overall classification accuracy indices ranged from 0.680 to 0.781, and the overall classification consistency indices ranged from 0.569 to 0.695 across grades. Grade 5 had the lowest overall classification accuracy and consistency indices.

For the Comprehension composite, as shown in Table 5.7.3.1, the overall classification accuracy indices ranged from 0.640 to 0.702, and the overall classification consistency indices ranged from 0.529 to 0.603 across grades. Grade 1 had the lowest overall classification accuracy and consistency indices.

For the Overall composite, as shown in Table 5.7.4.1, the overall classification accuracy indices ranged from 0.735 to 0.824, and the overall classification consistency indices ranged from 0.640 to 0.754 across grades. Grade 5 had the lowest overall classification accuracy and consistency indices.

The results reveal that Grade 5 had the lowest overall classification accuracy and consistency indices for three out of the four composites (Oral, Literacy, and Overall), while Grade 1 had the lowest overall classification accuracy and consistency indices for the Comprehension composite.

From an accountability perspective, the most important indices for test users and policy makers to examine are the marginal classification accuracy and consistency indices. We report for each composite the range of the marginal classification accuracy and consistency indices for the composite scale scores across grades and then highlight the grade (and the cut point within that grade) that had the lowest marginal classification accuracy and the lowest consistency indices.

For the Oral composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.863 to 0.998 (Table 5.7.1.2), and the marginal classification consistency indices ranged from 0.808 to 0.997 (Table 5.7.1.3). Grade 5, at the PL 4/5 cut point, had the lowest marginal classification accuracy and consistency indices. Note that Grade 5 also had the lowest overall classification accuracy and consistency indices for the Oral composite. The low marginal classification accuracy and consistency at the PL 4/5 cut point appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal classification accuracy and consistency indices for the Grade 5 Oral composite are still in the range of 0.80 and 0.90.

For the Literacy composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.851 to 0.999 (Table 5.7.2.2), and the marginal classification consistency indices ranged from 0.793 to 0.998 (Table 5.7.2.3). Grade 5, at the PL 3/4 cut point, had the lowest marginal classification accuracy and consistency indices. Note that the overall classification accuracy and consistency indices for the Literacy composite were the second lowest. The low marginal classification accuracy and consistency at the PL 3/4 cut point appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal accuracy and consistency indices for the Grades 5 Literacy composite are still in the 0.70 to 0.90 range.

For the Comprehension composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.894 to 0.982 (Table 5.7.3.2), and the marginal classification consistency indices ranged from 0.852 to 0.975 (Table 5.7.3.3). Grade 1, at the PL 2/3 cut point, had the lowest marginal classification accuracy and consistency indices. Note that Grade 1 also had the lowest overall classification accuracy and consistency indices for the Comprehension composite. The low marginal classification accuracy and consistency at the PL 2/3 cut point appeared to have contributed to its low overall classification accuracy and

consistency. However, it should be noted that the marginal accuracy and consistency indices for the Grade 1 Comprehension composite are still in the high 0.80 to mid-0.90 range.

For the Overall composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.902 to 0.993 (Table 5.7.4.2), and the marginal classification consistency indices ranged from 0.867 to 0.993 (Table 5.7.4.3). Grade 5 had the lowest marginal classification accuracy at the PL 4/5 cut point, and consistency indices at the PL 3/4 cut point. Note that Grade 5 also had the lowest overall classification accuracy and consistency indices for the Overall composite. The low marginal classification accuracy and consistency at the PL 4/5 and PL 3/4 cut points, respectively appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal accuracy and consistency indices for the Grade 5 Overall composite are still in the 0.80 to 0.90 range.

When we compared the overall and marginal classification accuracy and consistency indices for the composites for a particular grade, we saw that in many instances they told the same story (i.e., for a given grade, if the overall classification accuracy and consistency indices were low, then the marginal classification accuracy and consistency indices also tended to be low). This was especially true for Grade 5 for three of the four composites (Oral, Literacy, and Overall). Grade 5 had the lowest overall and marginal classification accuracy and consistency indices for these composites. Similarly, Grade 1 had the lowest overall and marginal classification accuracy and consistency indices for the Comprehension composite. In addition, the lowest marginal classification accuracy and consistency based on the composite scale scores occurred at the PL 2/PL 3, PL 3/PL 4, and PL 4/PL 5 cut points. A higher number of proficiency levels typically results in cut points that are closer to each other than if there were a smaller number of proficiency levels. We would expect marginal classification accuracy and consistency to vary for different ability levels due to variation in measurement accuracy. That is, the further away the students' composite scale scores are from the cut points, the smaller the classification errors would be, or the more accurate the classification decisions would be. With many proficiency levels, there are more student composite scale scores near the cut points than there would be if there were fewer with only two proficiency levels. Therefore, the higher the number of proficiency levels, the higher the probability that students would be misclassified (Ercikan & Julian, 2002). The marginal classification accuracy and consistency indices based on the

composite scale scores for cut points that are in the middle range tend to be lower than for other cut points, as we might expect.

Assessment experts have issued little guidance to aid in making judgments about the ideal or expected levels of decision consistency and accuracy needed for educational assessments that report composite scale scores. From an accountability perspective, the most important indices are the marginal classification accuracy and consistency indices. The marginal classification accuracy and consistency indices were at or above 0.793 for all four composites. Additionally, the marginal classification accuracy and consistency indices were at or above 0.867 for the Overall composite scale score, which is the primary score that WIDA Consortium states use when making accountability decisions.

5.7.1 Oral

Table 5.7.1.1

Overall Accuracy and Consistency of Classification Indices: Oral S601 Online

Grade	Accuracy	Consistency
1	0.702	0.597
2	0.708	0.606
3	0.677	0.579
4	0.646	0.545
5	0.644	0.532
6	0.731	0.629
7	0.719	0.614
8	0.710	0.605
9	0.753	0.660
10	0.752	0.660
11	0.747	0.654
12	0.755	0.664

Table 5.7.1.2

Marginal Classification Accuracy Indices Based on the Composite Scale Scores at the Cut Points: Oral S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.951	0.921	0.900	0.936	0.991
2	0.957	0.917	0.902	0.937	0.992
3	0.961	0.922	0.889	0.906	0.990
4	0.983	0.964	0.922	0.883	0.892
5	0.977	0.958	0.916	0.863	0.926
6	0.970	0.933	0.898	0.938	0.990
7	0.962	0.927	0.897	0.943	0.989
8	0.954	0.922	0.897	0.944	0.989
9	0.939	0.916	0.920	0.978	0.998
10	0.936	0.915	0.922	0.978	0.997
11	0.938	0.913	0.921	0.975	0.997
12	0.935	0.910	0.926	0.983	N/A

Table 5.7.1.3

Marginal Classification Consistency Indices Based on the Composite Scale Scores at the Cut Points: Oral S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.932	0.888	0.861	0.908	0.989
2	0.939	0.883	0.862	0.914	0.991
3	0.944	0.889	0.845	0.887	0.989
4	0.977	0.948	0.890	0.835	0.871
5	0.968	0.938	0.883	0.808	0.899
6	0.958	0.904	0.858	0.913	0.989
7	0.946	0.896	0.856	0.918	0.987
8	0.936	0.889	0.856	0.920	0.987
9	0.914	0.881	0.887	0.971	0.997
10	0.910	0.880	0.890	0.971	0.997
11	0.912	0.877	0.887	0.969	0.997
12	0.908	0.873	0.893	0.979	N/A

5.7.2 Literacy

Table 5.7.2.1

Overall Accuracy and Consistency of Classification Indices: Litr S601 Online

Grade	Accuracy	Consistency
1	0.775	0.690
2	0.760	0.667
3	0.735	0.635
4	0.687	0.577
5	0.680	0.569
6	0.781	0.695
7	0.770	0.679
8	0.757	0.663
9	0.738	0.637
10	0.744	0.645
11	0.742	0.643
12	0.750	0.653

Table 5.7.2.2

Marginal Classification Accuracy Indices Based on the Composite Scale Scores at the Cut Points: Litr S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.896	0.910	0.976	0.994	N/A
2	0.943	0.891	0.937	0.990	N/A
3	0.949	0.905	0.904	0.977	0.999
4	0.958	0.923	0.865	0.938	0.989
5	0.958	0.926	0.851	0.940	0.992
6	0.944	0.903	0.937	0.997	N/A
7	0.941	0.902	0.933	0.995	N/A
8	0.935	0.900	0.927	0.994	N/A
9	0.947	0.903	0.912	0.976	0.998
10	0.944	0.898	0.921	0.980	N/A
11	0.944	0.897	0.922	0.980	N/A
12	0.933	0.891	0.938	0.987	N/A

Table 5.7.2.3

Marginal Classification Consistency Indices Based on the Composite Scale Scores at the Cut Points: Litr S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.853	0.871	0.967	0.993	N/A
2	0.917	0.848	0.910	0.987	N/A
3	0.927	0.867	0.865	0.967	0.998
4	0.940	0.889	0.811	0.915	0.988
5	0.940	0.895	0.793	0.913	0.991
6	0.922	0.863	0.910	0.996	N/A
7	0.917	0.861	0.903	0.993	N/A
8	0.909	0.859	0.896	0.992	N/A
9	0.926	0.862	0.875	0.967	0.998
10	0.921	0.857	0.888	0.972	N/A
11	0.921	0.854	0.890	0.971	N/A
12	0.906	0.848	0.912	0.982	N/A

5.7.3 Comprehension

Table 5.7.3.1

Overall Accuracy and Consistency of Classification Indices: Cphn S601 Online

Grade	Accuracy	Consistency
1	0.640	0.529
2	0.691	0.585
3	0.669	0.567
4	0.697	0.600
5	0.677	0.579
6	0.702	0.602
7	0.687	0.585
8	0.682	0.580
9	0.702	0.603
10	0.697	0.599
11	0.693	0.594
12	0.701	0.602

Table 5.7.3.2

Marginal Classification Accuracy Indices Based on the Composite Scale Scores at the Cut Points: Cphn S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.923	0.894	0.908	0.931	0.968
2	0.961	0.914	0.915	0.933	0.963
3	0.944	0.916	0.916	0.925	0.951
4	0.982	0.949	0.925	0.909	0.920
5	0.970	0.946	0.920	0.906	0.921
6	0.962	0.922	0.913	0.929	0.969
7	0.953	0.923	0.914	0.927	0.962
8	0.945	0.924	0.917	0.927	0.959
9	0.952	0.926	0.921	0.931	0.965
10	0.949	0.926	0.921	0.931	0.962
11	0.949	0.925	0.919	0.930	0.962
12	0.943	0.923	0.922	0.939	0.969

Table 5.7.3.3

Marginal Classification Consistency Indices Based on the Composite Scale Scores at the Cut Points: Cphn S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.894	0.852	0.873	0.902	0.954
2	0.944	0.878	0.880	0.905	0.948
3	0.921	0.881	0.882	0.895	0.930
4	0.975	0.929	0.893	0.874	0.887
5	0.959	0.923	0.887	0.870	0.889
6	0.947	0.890	0.879	0.901	0.955
7	0.934	0.892	0.880	0.898	0.945
8	0.923	0.893	0.885	0.898	0.940
9	0.934	0.896	0.889	0.904	0.948
10	0.930	0.896	0.890	0.904	0.945
11	0.929	0.894	0.887	0.902	0.944
12	0.920	0.891	0.892	0.914	0.955

5.7.4 Overall

Table 5.7.4.1

Overall Accuracy and Consistency of Classification Indices: Over S601 Online

Grade	Accuracy	Consistency
1	0.814	0.740
2	0.814	0.740
3	0.786	0.705
4	0.737	0.645
5	0.735	0.640
6	0.824	0.754
7	0.813	0.739
8	0.806	0.730
9	0.806	0.730
10	0.810	0.733
11	0.807	0.731
12	0.814	0.739

Table 5.7.4.2

Marginal Classification Accuracy Indices Based on the Composite Scale Scores at the Cut Points: Over S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.936	0.916	0.970	0.992	N/A
2	0.961	0.923	0.944	0.987	N/A
3	0.965	0.932	0.918	0.971	N/A
4	0.978	0.954	0.910	0.907	0.985
5	0.975	0.953	0.906	0.902	0.991
6	0.969	0.934	0.928	0.993	N/A
7	0.963	0.933	0.927	0.991	N/A
8	0.957	0.930	0.931	0.989	N/A
9	0.958	0.930	0.935	0.983	N/A
10	0.954	0.927	0.942	0.987	N/A
11	0.955	0.927	0.941	0.985	N/A
12	0.948	0.921	0.953	0.992	N/A

Table 5.7.4.3

Marginal Classification Consistency Indices Based on the Composite Scale Scores at the Cut Points: Over S601 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.909	0.881	0.958	0.992	N/A
2	0.944	0.891	0.920	0.985	N/A
3	0.951	0.904	0.884	0.965	N/A
4	0.968	0.935	0.874	0.874	0.984
5	0.964	0.933	0.867	0.874	0.991
6	0.956	0.907	0.898	0.993	N/A
7	0.948	0.904	0.896	0.990	N/A
8	0.940	0.900	0.901	0.987	N/A
9	0.941	0.901	0.907	0.979	N/A
10	0.936	0.897	0.917	0.983	N/A
11	0.937	0.896	0.916	0.981	N/A
12	0.927	0.889	0.933	0.990	N/A

6. Quality Control

6.1 Content Development Quality Control

The Center for Applied Linguistics (CAL) utilizes educators and other consultants at a number of phases throughout the test development cycle. These educators and consultants are recruited, vetted, and trained by CAL and/or WIDA and make crucial contributions to these phases of the test development cycle. The phases of development in which educators or consultants are involved, as well as the procedures and criteria for recruitment and training, are described below.

Theme Generation

During theme generation, CAL and WIDA recruit educators to generate raw ideas to be used in new item development. Educators with ESL or content-area expertise and two or more years of teaching experience in a WIDA state (in the grade cluster for which they will generate themes) are invited to participate. Recruitment also focuses on a geographical distribution of educators from across the consortium. Upon selection, educators participate in a short training that introduces the theme generation process, along with how to understand the item specifications that they use to generate themes.

Item Writing

CAL recruits professional item writers to generate raw item/task content based on the ideas from theme generation. To recruit item writers, CAL has a standing announcement on its website asking prospective item writers to submit their resumes and fill out a survey describing their past item writing experience. CAL selects individuals with significant experience in writing items, both in large-scale assessment programs (ESL/EFL or ELA) and in other contexts (e.g., writing items for assessment programs in university-based ESL programs).

Item writers undergo a 90-minute orientation prior to beginning item writing. This training focuses on the item specifications, the process and procedures, the item writing checklist, the acceptance criteria for the items, and the security protocols. Item writers also receive an item writing handbook, which formalizes the content of the orientation, along with assignment of themes to develop and the associated item specifications. After the orientation, CAL Language Testing Specialists and managers provide feedback to the item writers on the items, focusing on alignment with the item writing checklist and the item specifications. After completion of item

writing for a given development cycle, item writers are evaluated by CAL staff for their compliance with the requirements and the quality of their items.

Standards Expert Review

After items have been drafted by item writers, CAL Language Testing Specialists review all of the raw content internally. This review focuses on determining which sets of items will move on to further development and which will be discontinued, based on criteria from an item review checklist. The Language Testing Specialists then do minor editing and formatting to the items to make sure that they are complete, with no stray comments or other editorial notes from previous drafts, and they produce a short questionnaire for each set of items that becomes part of Standards Expert review. The purpose of Standards Expert review is to ensure that the items are appropriate for the grade level and intended difficulty level in terms of both the content and the language, and the items have not drifted from their intended target between theme generation and item writing. The questionnaires produced by CAL's Language Testing Specialists guide the Standards Experts through the review process, asking questions specific to the purpose of this review.

Educators are recruited jointly by CAL and WIDA to serve as Standards Experts; educators with ESL or content-area expertise and two or more years of teaching experience in a WIDA state are invited to participate. Recruitment also focuses on a geographical distribution of educators from across the consortium. Standards Experts receive written instructions and a questionnaire to complete for each set of items they review.

Bias and Sensitivity and Content Review

After Standards Expert Review has been completed, all items undergo an additional phase of review and revision internal to CAL, leading up to Bias & Sensitivity and Content Review. These are technically two separate reviews, although a single recruitment effort is conducted by WIDA, and the reviews occur consecutively in a single week (generally 3 days for Content review followed by 2 days for Bias & Sensitivity review). As with other reviews, educators for Content review must have at least 2 years of ESL teaching experience (with a preference for content-area experience as well). Recruitment also focuses on selecting educators with a variety of cultural and linguistic backgrounds and obtaining a geographical distribution of educators from across the consortium. Recruitment for Bias & Sensitivity review focuses on selecting

educators with culturally and linguistically diverse backgrounds who have experience interacting with English learners from a range of cultural, regional, religious, linguistic, ethnic, and socioeconomic backgrounds.

At the beginning of both Bias & Sensitivity and Content review meetings, CAL and WIDA staff conduct an intensive training to orient the reviewers to the specific purpose of the review (Bias & Sensitivity or Content), how to use the review checklist and what to look for in the review, and the procedures and security protocols for the review. Then, the reviews are conducted in breakout groups by grade cluster (or combinations of grade clusters; for example, Bias & Sensitivity review of Grade 1 and Grades 2–3 is often combined). Although Bias & Sensitivity and Content reviews are generally held in-person, the reviews for the Writing domain occur virtually each year due to timeline constraints. For both the in-person and virtual contexts, CAL and WIDA facilitators are present in each breakout group to guide the educators in their reviews of the materials.

Writing Tryouts

All tasks in the Writing domain are subject to tryouts in the field. The Writing tryouts only occur once the tasks have been through a thorough Bias & Sensitivity and Content review and subsequent revision. CAL and WIDA recruit educators who are willing to administer the Writing tasks to their students; these educators are classroom ESL or content teachers who work with ELs. All students who participate are required to have parent/guardian consent.

Once the students complete the Writing tasks, both the students and educators fill out questionnaires. Student questionnaires focus on whether the students understood the task, their engagement with the task, and their ability to complete the task; educator surveys ask the teachers to evaluate the effectiveness of the task input, the appropriateness of the task, the comparability of the task with other classroom-based writing tasks, and the ability of the students to complete the task.

CAL provides the teachers with a number of documents outlining the procedures for administering the tasks, recording student responses to the tasks, recording student and teacher responses to the questionnaires, and protecting the personally identifiable information of the students. CAL staff are also available throughout the tryouts process to answer any questions the teachers might have. Following the Writing tryouts, CAL specialists review the writing

responses both qualitatively and quantitatively, providing WIDA with a report on how the Writing tasks performed.

6.2 Test Administration Quality Control

This section describes how WIDA monitors test administration to ensure standardized test administration procedures are implemented with fidelity across districts and schools. To support standardized administrations, WIDA provides test administrators with a series of resources, such as a Test Administration Manual, a training course, and a Test Administration script for each assessment.

Qualifications of Test Administrators

Before, during, and after a state's testing window, educators hold various roles to ensure all tasks are carried out for successful test administration. These roles include Test Coordinators at the district and school level and Test Administrators. The Test Administrator administers and monitors the test, and is also responsible for managing student data prior to, during, and after testing.

WIDA has worked directly with each state education agency to develop the ACCESS for ELLs Checklist for the school year. This list highlights all tasks that need to be completed before, during, and after testing within a school or district and outlines which tasks are assigned to Test Coordinators at the district and school level and to Test Administrators. It also provides additional guidance that a state expects test administrators to follow as they prepare for and administer the ACCESS for ELLs suite of assessments.

Test Administrators are responsible for reviewing each state's checklist in detail prior to completing any training and for working with the district or school Test Coordinator to complete these tasks. The state's checklist can be found in the training course and on each state's WIDA webpage at www.wida.us/membership/states.

The training course within the WIDA Secure Portal (https://www.wida.us/login.aspx) is 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 each state's testing window. WIDA user accounts provide access to the training course and Facilitator Toolkit within the WIDA Secure Portal. Educators must pass an administration quiz at

the end of the training with a score of 80% or higher. WIDA recommends taking the quiz immediately after completing the training. There is no limit to the number of times educators can attempt the quiz. Once individuals pass an administration quiz, training certificates within the WIDA Secure Portal are updated to reflect their status as a certified Test Administrator for that component of the assessment suite.

Paper Testing (for Writing Grades 1–3)

Depending on state, district, and school policy, not all Test Administrators will be responsible for initially labeling and/or bubbling booklets. However, it is the responsibility of all Test Administrators and Test Coordinators to ensure that correct and complete information is either labeled or bubbled in each student booklet. Each state's ACCESS for ELLs Checklist has more information on who is responsible for each task related to materials management in the state.

To ensure all booklets have the detailed and necessary information needed to score, all Test Administrators must adhere to the following:

- Prior to administration
    - Review labels and/or bubbled information to ensure all student information is accurate.
    - Complete labeling or bubbling if needed.
- During administration
    - Distribute the test booklets, as applicable, to the correct students.
    - Verify that students have been given their assigned booklet.
- Immediately following administration
    - Collect all material from all students.
    - Review student test booklets once more for any errors or discrepancies in student information.
    - Confirm all necessary fields are completed and all necessary labels are correctly adhered to student test booklets.
    - Ensure all booklets are in proper condition to be returned, with no loose or damaged pages.

- Return test materials to a Test Coordinator, or store the booklets in a secure area until they can be handed over to a Test Coordinator.

Failure to address incorrect, missing, or incomplete booklet information and labels may result in late reporting or no student score. In addition, the WIDA Consortium's national research agenda relies on complete and accurate student demographic data to inform the field and benefit English language learners.

When preparing test materials for return to DRC, test administrators need to confirm that any booklet that contains student response information has either a Pre-ID Label or a District/School Label with bubbled student information. If a booklet is unused, there is no need to place any labels on the booklet. Placing a label on a booklet will cause it to be processed (and either scored, if the label is a Pre-ID or School/District label, or not scored, if it is a Do Not Process label).

6.3 Rater Quality Control

Rater Training

Students who take the ACCESS for ELLs Paper Speaking test have their spoken responses scored by the Test Administrator who administered the Speaking test. Another term for this Test Administrator is *rater*. Raters must be trained and certified, so we can be confident that they interpret students' spoken language consistently and fairly and that the scores are reported according to the WIDA English language proficiency standards. WIDA provides several different types of resources to support raters' training and reliability.

Students who take ACCESS for ELLs Online have their spoken responses digitally recorded and then scored centrally by DRC's trained raters. Students who take ACCESS for ELLs Paper have their spoken responses scored in real time by the Test Administrator who administers the Speaking test. In both cases, it is important that the individual who scores the spoken responses is trained and certified.

WIDA provides a series of training modules in the Secure Portal on the WIDA website. ACCESS for ELLs Speaking test raters should complete three core modules:

1. Overview and Test Structure

2. Speaking Assessment Scoring Practice
3. Speaking Assessment Recommended Practice

WIDA strongly recommends that all new raters complete all three of these modules. These modules provide a comprehensive introduction to the ACCESS for ELLs Speaking test and the opportunity to learn how to score students' spoken English reliably using the ACCESS for ELLs Speaking Scoring Scale.

In addition to the modules described above, WIDA also releases supplemental training materials each year to refamiliarize experienced raters with the Speaking Scoring Scale and introduce new Speaking tasks and sample responses for the coming year. These materials, called Supplemental Training for the Speaking Assessment, reflect the Speaking tasks that will appear on the test in the current year. WIDA recommends that all raters (new and experienced) engage with these supplementary materials at the start of each scoring season. Reading and reviewing these materials will help raters maintain their reliability from year to year and contribute to the fairness of test scores awarded to all students.

Rater Certification

After completing the training modules described in the section above, new raters should take the relevant certification quiz. WIDA provides two quizzes: one for raters who will evaluate students in Grades 1–5 and another for raters who will evaluate students in Grades 6–12. Raters should take the appropriate quiz.

The purpose of the quiz is to ensure that raters have internalized the Speaking Scoring Scale and can apply it consistently. Only raters who pass the quiz(zes) should administer and score the ACCESS for ELLs Paper Speaking test.

Checklist for Rater Training, Monitoring, and Recertification

- ✓ New raters complete all Speaking Assessment Training
- ✓ New raters take and pass the appropriate certification quizzes
- ✓ All raters recertify at the start of each testing season (review new materials, retake quiz)

- ✓ Only certified raters administer and score the ACCESS for ELLs Speaking test
- ✓ Raters do not evaluate their own students, if at all possible
- ✓ Rater reliability and/or score point distributions are monitored regularly

For more information on Writing rater QC, please refer to section 3.2.2.

6.4 Score Reporting Quality Control

WIDA conducts an annual score reporting quality control process to (1) verify the accuracy of paper-based test scores (i.e., ACCESS for ELLs Paper, Kindergarten ACCESS for ELLs, and Alternate ACCESS) and (2) verify the accuracy of all score reports (the Individual Student Report, the Student Roster Report, the School Frequency Report, the District Frequency Report, and the State Frequency Report) for both ACCESS (Online, Paper, and Kindergarten) and Alternate ACCESS.

The Score Reporting quality control is conducted at DRC's offices in Maple Grove, Minnesota. The team generally includes five state education agency representatives, one CAL employee, and four WIDA employees. This team examines data from three districts: a primary district, for quality control of all score reports; a secondary district, for quality control of State Frequency Reports only; and a tertiary district for quality control of paper-based tests only.

After an introductory presentation, which includes details of the quality control processes undertaken by DRC and WIDA and instructions on using the data entry tools, panelists begin by confirming the scoring of ACCESS Paper. Using the information in the State Student Response file, panelists enter the grade level, grade level cluster, tier, the Listening and Reading responses, and the Speaking and Writing scores into the data entry tool. The tool then calculates the student's raw scores and, using a series of look-ups, the student's scale score, proficiency level score, and confidence bands for all domains and composites. Panelists check student scores on the Individual Student Reports against those calculations. Any discrepancies are brought to the attention of the WIDA facilitator who investigates and, if there seems to be an issue with the report (rather than the data entry or data entry tool), discusses the issue further with DRC.

The panelists follow a similar process with the Kindergarten ACCESS tests, but with the raw scores for these tests copied directly from the response booklets.

After checking the paper-based tests, panelists turn their attention to the score reports. Panelists first check both the demographic information and the student scores in the Individual Student Reports against the information in the Student Roster Reports. Again, any discrepancies are brought to the attention of the facilitator, who investigates and discusses the issue with DRC if necessary. Panelists use the verified Individual Student Reports to check the Student Roster Report. Once the Student Roster Report is verified, panelists use it to check the State Frequency Report; they then use the verified State Frequency Reports to check the District Frequency Report. Finally, panelists check the State Frequency Reports against verified District Frequency Reports from the primary district along with District Frequency Reports from the secondary district.

6.5 Data Forensic Quality Control

Incidence of student plagiarism

DRC and WIDA have identified and confirmed instances of a student/students plagiarizing responses of the Speaking and/or Writing tests for mostly clusters 68 and 912 items. While scoring student responses, DRC identified these students' responses as not being authentic to the student. WIDA staff have confirmed that students accessed the internet to look up specific wording from the task and to use information from a website in order to respond to the task. Some students produced spoken responses by utilizing an artificial voice (not the student's own voice), via either translation software or screen reading functionality. When plagiarism was identified, the SEA representative in the state where the infraction occurred was notified immediately, and WIDA requested direction about those students' scores. All responses containing plagiarized content will receive a nonscorable code of "Invalid Indecipherable." This impacted 765 students in Speaking and 432 students in Writing across 40 states/territories. Below is the summary of the number of students suspected of plagiarism in Speaking and Writing domains by state.

Table 6.5.1 Number of Instances of Plagiarism

State	Student Counts	
	Speaking	Writing
AK*	5	0
AL*	1	2
BIE*	1	1
CNMI*	2	3
CO*	19	3
DC	1	0
DE*	2	2
DODEA*	1	0
FL*	0	16
GA*	25	24
HI*	2	3
ID*	2	2
IL*	103	52
IN*	21	8
KY*	12	2
MA*	26	14
ME*	4	2
MD*	29	32
MI*	45	26
MN*	16	15
MO*	27	8

State	Student Counts	
	Speaking	Writing
MT*	1	3
NC*	54	27
ND*	0	2
NH*	1	0
NJ*	29	14
NM*	30	7
NV*	37	17
OK*	55	11
PA*	27	31
RI*	13	16
SC*	16	8
SD*	3	0
TN*	19	8
UT*	22	12
VA*	41	29
VI*	0	1
VT*	1	0
WA*	38	22
WI*	34	9
State Totals	**Student Totals**	
40	**765**	**432**

* = states where scoring is complete and all flagged suspected instances of plagiarism have been reported to SEA. Counts represent # of students that were flagged for suspected plagiarism. Some students were flagged for multiple responses, so overall response count flagged is higher.

Suspected Item Exposure stopped here

Beginning in February 2023, WIDA, state partners, and Caveon identified 87 posts on social media containing ACCESS related content, out of which 59 posts are related to sample items or practice materials. A total of 13 items were identified from the remaining 28 posts. All posts were removed from social media upon request. A memo was distributed by WIDA to all SEAs on March 6, 2023, to report the issue and no posts were identified after March 9, 2023.

An item is suspected of being exposed if any content appears on social media. The WIDA test development team reviewed images and videos to identify the exact screens that clearly contained content related to tasks, prompts, and response options. The WIDA psychometrics team conducted analyses comparing item performance before and after items were exposed against overall item performance. Item parameters from the last year were compared against this year's item parameters using the data with potential item exposure. Item statistics were also reviewed and compared week-to- week before items were removed from social media. Given that these posts were promptly removed from local devices or social media, the results suggested little variation regarding item performance. WIDA has decided to retain operational items for scoring, but exposed items were excluded from item calibration for verification studies for operational items and will not appear on the 2023-24 test administration. Any field test items that were exposed will not be part of next year's operational test.

Table 6.5.2 provides the summary of these 13 items flagged for exposure in Speaking, Writing, and Listening tests from the 2022-2023 administration. Four speaking items were from the G6-8 cluster, and the remaining nine items were all from the G9-12 cluster across three domains. Nine items were scored as operational items and four items were unscored field test items. The displacement index represents differences of item parameters between last year and this year. Displacement values between the two points of item parameters did not show a greater difference in logit values representing item difficulty, which suggested items did not get significantly easier or harder due to exposure.

Table 6.5.2 Number of Item Exposures

Domain	Item Status	Number of Items	Cluster
Speaking	Operational	3	68
Speaking	Operational	4	912
Speaking	field Test	1	68
Writing	Operational	1	912
Writing	field Test	1	912
Listening	Operational	1	912
Listening	field Test	2	912

Caveon Data Forensic Analysis Results

WIDA hired Caveon to perform data forensic analysis during the 2022–2023 test administration cycle to examine whether ACCESS data has been compromised or has evidence of item exposure.

Caveon security statistics are based on mathematical models, where the test response data are used to create a baseline model of normal or "typical" test taking among that population. Individuals or groups are then compared to the baseline, and observations that are significantly different from the baseline are flagged as anomalous. Caveon's statistics are designed to be robust but also conservative regarding which and how many individuals or groups are flagged as anomalous, thereby reducing the chances of false-positive detections.

Data forensics analysis was performed after the administration window for the following administrations:

- December 2022 through August 2023 online multistage adaptive test administrations, Listening and Reading domains
- December 2022 through August 2023 paper fixed-form administrations, Listening and Reading domains

The analysis utilized several of Caveon's security statistics to detect evidence of whether the assessment instrument has been compromised through disclosure of the content. This analysis attempted to understand where and when disclosure of the test content may have occurred and what items and forms may have been affected. Results of this analysis might enable WIDA to take specific actions to limit the impact of disclosed content. Such actions may include:

- Republishing or reworking items or forms
- Rotating disclosed items to limit their exposure
- Designing a republication or rotation strategy for future items and forms

Caveon security statistics were computed for each individual test instance. These data were aggregated or summarized at the group level. The aggregated statistics were compared against the population model.

Analysis of Tests

Caveon aggregated the data according to individual test forms using the security statistics to determine whether rates of detections by the security statistics were higher for certain test forms. For fixed-form paper tests, two forms—A and B/C—were analyzed. For the multistage adaptive test, there is a finite number of ways a student could progress through the test. Caveon analyzed each pathway as a separate form. Higher rates of security detections for a specific form of the test suggest that compromise of the form may have occurred.

Analysis of Items

Item security: In this portion of the analysis, the security of the items was evaluated using aberrance statistics. Aberrance statistics detect test-taking behaviors such as answering difficult items correctly but answering easy items incorrectly, or unusual patterns in the time taken to answer test items. In the absence of security issues, aberrant test taking is expected to be the result of poor or uneven test preparation, illness or other physical malady, mental and emotional distractions, and so forth. These factors usually result in lower levels of test performance. When aberrance is associated with higher performance, however, test fraud may have occurred, such as pre-knowledge of test content. By applying aberrance measures and comparing the performance between aberrant and non-aberrant test instances on individual items, inferences can be made about item security.

Item performance changes: Analysis of item performance changes tracks individual item performance rates over time. The item performance shifts are measured within the context of the item response theory model and adjusted for varying test-taker performance levels. This means that detected performance shifts are invariant to fluctuations in the test-taker population. When performance shifts indicate the item has become significantly easier, the item may have been disclosed. Items with significant performance shifts become candidates for revision or replacement. Item performance shifts were detected with a granularity of 1 week, where Monday to Sunday represents 1 week.

Analysis of Groups

Analysis by week: This analysis aggregates the data according to the week in which the test was taken to identify whether security threats and pass rates appeared to be more prevalent at certain times during the testing window. Increases in scores or security detections during certain periods of time suggest the content may have been disclosed at some point prior to that time. This analysis also includes a form-date grouping to determine if increasing security threats are associated with a particular form of the test. This analysis is performed for online and paper tests, where relevant test date data are provided.

Analysis of WIDA jurisdictions: Caveon analyzed WIDA member jurisdictions (states and districts) to determine whether rates of detections by the security statistics were higher for certain jurisdictions. This analysis is intended to detect whether compromise at the state or member jurisdiction level potentially occurred. This analysis is performed for online and paper tests.

Analysis of administration mode: Caveon aggregates the data according to administration mode (i.e., online versus paper) to determine if security threats are associated with the mode of testing.

Other Analyses

Analysis of mean score over time was used to identify whether mean scores increased over time during the testing window. Increases in scores over time suggest the content may have been disclosed during the testing window.

Findings of Data Forensic Analyses

Generally, no major data forensic anomalies were observed across WIDA states. A few minor localized anomalies associated with items are under WIDA's investigation.

References

ACCESS Test Practice and Sample Items | WIDA. (n.d.). Wida.wisc.edu; Wisconsin Center for Education Research at the University of Wisconsin–Madison. https://wida.wisc.edu/assess/access/preparing-students/practice

Allen, N. L., Carlson, J. E., & Zalanak, C. A. (1999). *The NAEP 1996 technical report.* National Center for Education Statistics.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.

American Institutes of Research. (2018). *ELPA21 technical report, part I—Summative assessment*. Author.

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

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R.* Springer International Publishing AG.

Brennan, R. (2004). *Linking with equivalent group or single group design (LEGS)* (Version 2.0) [Computer software]. Center for Advanced Studies in Measurement and Assessment.

Chapman, M., Montee, M., & Musser, S. (2022). *ACCESS speaking educator perceptions and instructional practices study* (Unpublished WIDA Internal Report). WIDA, University of Wisconsin–Madison.

Center for Applied Linguistics. (2016). *ACCESS for ELLs Series 400 listening and reading scale maintenance: Technical brief.* Author.

Center for Applied Linguistics. (2017). *ACCESS for ELLs 2.0 speaking and writing score scale reconstruction: Technical brief.* Author.

Center for Applied Linguistics. (2019). *Maintaining the ACCESS for ELLs online writing scale: Preparations for the Series 501 redesign: Technical brief*. Author.

Cook, H. G., & MacGregor, D. (2017). *The ACCESS for ELLs 2.0 2016 standard-setting study* (Technical Report). Board of Regents of the University of Wisconsin System.

Crabtree, A. R. (2016). *Psychometric properties of technology-enhanced item formats: An evaluation of construct validity and technical characteristics*. Unpublished doctoral dissertation, University of Iowa. https://doi.org/10.17077/etd.922fbj4d

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

Department of Education, (2018). *A state's guide to the U.S. Department of Education's assessment peer review process.* U.S. Department of Education. Elementary and Secondary Education Act of 1965, amended 2015. 20 USC §6301-8961.

Engelhard, G., Jr., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments.* Routledge/Taylor & Francis Group.

Ercikan, K, & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269–294.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105–146). Macmillan.

Gottlieb, M. (2004). *English language proficiency standards for English language learners in kindergarten through grade 12: Framework for large-scale state and classroom assessment.* WIDA Consortium.

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17,* 221–240.

Kenyon, D. M. (2006). *Development and field test of ACCESS for ELLs* (WIDA Consortium Technical Report No. 1). Center for Applied Linguistics.

Kenyon, D. M., Ryu, J. R., & MacGregor, D. (2013). *Setting grade level cut scores for ACCESS for ELLs* (WIDA Consortium Technical Report No. 4). Center for Applied Linguistics.

Kim, A., Kondo, A., Blair, A., Mancilla, L., Chapman, & M., Wilmes, C. (2016). *Interpretation and use of K–12 language proficiency assessment score reports: Perspectives of educators and parents* (WCER Working Paper No. 2016-8).

Kim, A., S., Baghastani, S., Macgregor, D., & Ho, P. (2022). *Supporting K-12 educators' language assessment literacy via resources informed by validation* (Unpublished WIDA Internal Report). WIDA, University of Wisconsin–Madison.

Kim, A., S., Yumsek, M., Kemp, J., Chapman, M., & Cook, H. (2022). *Universal tools activation in English language proficiency assessments; A comparison of grades 1-12 English learner with and without disabilities* (Unpublished WIDA Internal Report). WIDA, University of Wisconsin–Madison.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement. *Journal of Educational Measurement, 29*, 285–307.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26,* 412–432.

Linacre, J. M. (1994). Sample size and item calibrations stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions, 13*(2), 696. http://www.rasch.org/rmt/rmt132i.htm

Linacre, J. M. (2002a). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2002b, Autumn). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). JAM Press.

Linacre, J. M. (2006). Winsteps Rasch analysis (Version 3.60.1) [Computer software]. http://www.winsteps.com

Linacre, J. M. (2020). *Reliability and separation of measures*. Winsteps. https://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (n.d.). *Displacement measures.* Winsteps. http://www.winsteps.com/winman/displacement.htm

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

MacGregor, D., & Ma, Y. (2022). *Speaking contrasting group study* (Unpublished WIDA Internal Report). WIDA, University of Wisconsin–Madison.

Mantel, N., & Haenszel, W. (1959). Statistical aspect of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Meyer, J. P. (2018). jMetrik [Computer software]. http://itemanalysis.com/jmetrik-download/

Min, S., & Bishop, K. (2022). *Evaluation of ACESS online multistage test design* (Unpublished WIDA Internal Report). WIDA, University of Wisconsin–Madison.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*(4), 151–363.

National Center on Educational Outcomes. (2021). *Universal design of assessments*. https://nceo.info/Assessments/universal_design#:~:text=Universal%20design%20principles%20include%20careful,of%20content%20and%20skills%20tested

Price, L. R., Lurie, A., Raju, N., Wilkins, C., & Zhu, J. (2006). Conditional standard errors of measurement for composite scores on the Wechsler Preschool and Primary Scale of Intelligence – Third edition. *Psychological Reports, 98*(1), 237–252.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson, & S. L. Hershberger (Eds.). *The new rules of measurement: What every psychologist and educator should know* (pp. 219–240). Psychology Press.

Rudner, L. (2001, Spring). Informed test component weighting. *Educational Measurement: Issues and Practice, 20*(1), 16–19.

Sahakyan, N., (2020). *Generating alternate overall composite scale scores for English Learners with disabilities who are missing domain scores in the ACCESS for ELLs assessment* (WIDA Technical Report).

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 329–347). Routledge.

Stahl, J. A., & Muckle, T. (2007). Investigating drift displacement in Rasch item calibrations. *Rasch Measurement Transactions, 21*(3), 1126–1127.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer, N. Dorans, D. Eignor, R. Flaugher, B. Green, R. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159–184). Lawrence Erlbaum Associates.

U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process.* https://oese.ed.gov/files/2020/07/assessmentpeerreview.pdf

WIDA Consortium. (2007). *English language proficiency standards and resource guide, 2007 edition, prekindergarten through grade 12*. Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2012). *2012 amplification of the English language development standards kindergarten–grade 12*. Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2022). *ACCESS for ELLs interpretive guide for score reports.* Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2021). *ACCESS for ELLs test administrator manual.* Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2021). *ACCESS for ELLs district and school test coordinator manual.* Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2021). *Test policy handbook.* Board of Regents of the University of Wisconsin System.

WIDA Consortium. (2020). *WIDA consortium English language proficiency assessment for grades 1-12 test and item design plan ACCESS for ELLs online annual summative and WIDA screener online.* Board of Regents of the University of Wisconsin System.

Wright, B.D. & Douglas, G.A. (1975). *Best test design and self-tailored testing*. Research memorandum, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* MESA Press.

Young, M. J., & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* (CSE Technical Report 475). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Erlbaum.

Zwick, R., & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In Y. Duanli, A. A. von Davier, & C. Lewis (Eds.), *Computer multistage testing: Theory and applications* (pp. 271–284). CRC Press.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R., Thayer, D. T., & Wingersky, M. (1993). *A simulation study of methods for addressing differential item functioning in computer-adaptive tests* (ETS Research Report RR-93-11). Educational Testing Service. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1993.tb01522.x

1. Acknowledgments

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

From CAL:

Tanya Bitterman, M.A.

Sofia Buitrago, MS.

Yage (Leah) Guo, Ph.D.

Seong Eun Hong, Ph.D.

Michele Kawood, M.S.Ed.

Justin Kelly, Ph.D.

Dorry M. Kenyon, Ph.D.

Reshmi Kumpakha, M.A.

Jung-Jung Lee, M.Sc.

Samantha Musser, M.A.

Aubrey Sahouria

Jasmine Tsai, M.Ed.

Frank Wucinski, M.A.

Shu Jing Yen, Ph.D.

Xin Yu, M.A.

From WIDA:

Kyoungwon Bishop, Ph.D.

Syed Hadi