



**Annual Technical Report for
ACCESS for ELLs
Online English Language Proficiency Test
Series 502, 2020–2021 Administration**

Annual Technical Report No. 17A

Prepared by:

Center for Applied Linguistics

Language Assessment Division
Psychometrics and Quantitative Research Team

March 2022

The WIDA ACCESS for ELLs Technical Advisory Committee

This report has been reviewed by the WIDA ACCESS for ELLs Technical Advisory Committee (TAC), which includes 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

Executive Summary

This is the 17th annual technical report on the ACCESS for ELLs English Language Proficiency test and the 5th report on the ACCESS for ELLs assessment delivered 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 2020–2021 ACCESS for ELLs 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 Series 502 Online was administered in the school year 2020–2021 in 35 states, the Bureau of Indian Education, the Department of Defense Education Activity, and the District of Columbia for a total of 38 state entities (henceforth “states”).

The ACCESS Series 502 Online data set used in this report included the results of 1,104,743 students as of September 2021. The final number of students who participated in the ACCESS Series 502 Online tests is 1,196,665. The grade with the most students was Grade 1, with 137,292, while the grade with the fewest students was Grade 12, with 36,576 students. Of the participating WIDA states, Georgia has the largest number of students, with 95,725, while the U.S. District of Columbia had the fewest, with 39 students.

During the 2020–2021 testing year, many states suspended in-person schooling due to the COVID-19 public health emergency. Based on a comparison with prior years' numbers of participating students, WIDA believes that 30% fewer students participated in ACCESS Series

502 testing than the ACCESS Series 501 testing. Further details on the impact of COVID-19 is contained in the ACCESS 2020–2021 *Year in Review Report*.

ACCESS for ELLs Series 502 was offered in two administrative formats, an online format (Grades 1–12) and a paper format (Kindergarten–Grade 12). The current report (WIDA ACCESS Technical Report 17A) provides technical information pertaining to ACCESS for ELLs Series 502 Online. A second report (WIDA ACCESS Technical Report 17B) provides technical information for the ACCESS for ELLs Series 502 Paper assessment, including the Kindergarten assessment.

Part 1:
Purpose, Design, Implementation

Contents

1. Purpose and Design of ACCESS	1-1
1.1. Purpose Statement	1-1
1.2. The WIDA Standards	1-1
1.3. The WIDA Proficiency Levels	1-3
1.4. Language Domains	1-4
1.5. Grade-Level Clusters	1-5
1.6. Tiers	1-5
2. Test Development	2-1
2.1. Item and Task Design	2-1
2.1.1. Listening Items	2-1
2.1.2. Reading Items	2-2
2.1.3. Writing Tasks	2-3
2.1.4. Speaking Tasks	2-3
2.2. Test Design	2-5
2.2.1. Listening	2-5
2.2.2. Reading	2-7
2.2.3. Writing	2-9
2.2.4. Speaking	2-11
2.3. Test Construction	2-13
2.3.1. Item Development	2-13
2.3.2. Field Testing	2-18
2.3.3. Item Review and Selection	2-24
3. Test Administration	3-1
3.1. Test Delivery	3-1
3.1.1. Listening and Reading	3-1
3.1.2. Writing	3-1
3.1.3. Speaking	3-1
3.2. Operational Administration	3-2
3.2.1. Administering the Test Practice	3-2
3.2.2. Listening Test Administration	3-2
3.2.3. Reading Test Administration	3-3
3.2.4. Writing Test Administration	3-3
3.2.5. Speaking Test Administration	3-4
3.2.6. Test Security	3-5
3.3. Fairness and Accessibility	3-5

3.3.1	Support Provided to All ELLs	3-6
3.3.2	Support Provided to ELLs with IEPs or 504 Plans.....	3-6
4.	Scoring	4-1
4.1.	Multiple Choice Scoring: Listening and Reading	4-1
4.2.	Scoring Performance-Based Tasks: Writing and Speaking.....	4-1
4.3.	Writing Scoring Scale	4-7
4.4.	Speaking Scoring Scale	4-10
5.	Summary of Score Reports	5-12
5.1.	Individual Student Report.....	5-12
5.2.	Other Reports.....	5-15

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 Online, described in this report, and ACCESS Paper, described in a companion report.

1.2. The WIDA Standards

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

- *Standard 1:* ELLs communicate in English for **Social and Instructional** purposes within the school setting.

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

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

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

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

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

The overarching goal of ACCESS for ELLs 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.

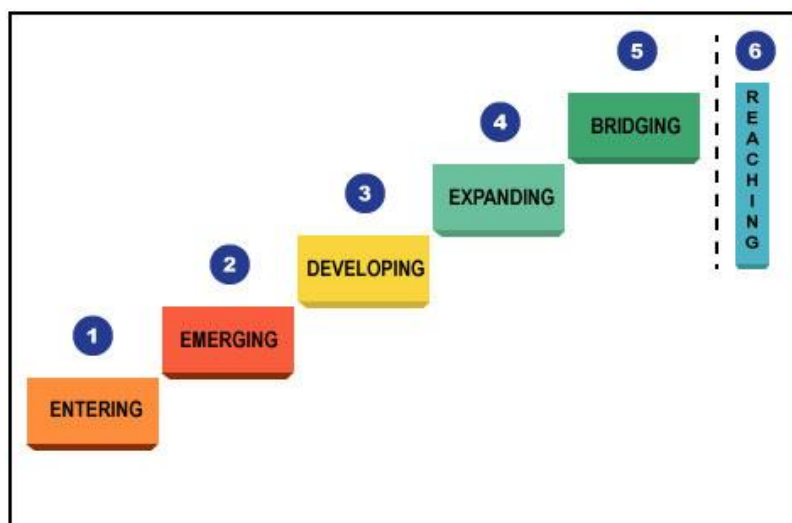


Figure 1. The language proficiency levels of the WIDA ELD Standards.

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: (a) vocabulary usage at the word/phrase level; (b) language forms and conventions at the sentence level; and (c) linguistic complexity at the discourse level. Vocabulary usage refers to students' increasing comprehension and production of the technical language required for success in the academic content areas. Language forms and conventions 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, that 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

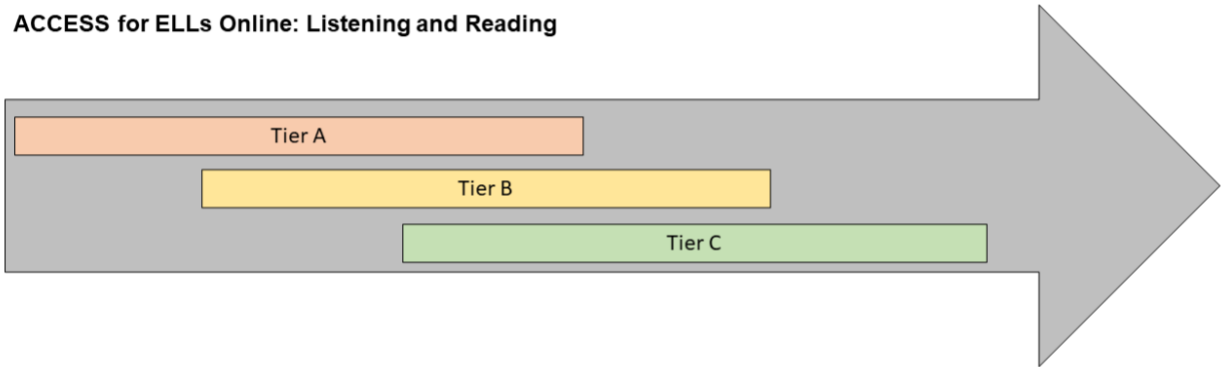
ACCESS is designed so that test paths or forms are appropriate to the proficiency level of individual students across the wide range of proficiencies described in the WIDA ELD Standards (Figure 2). Tests must be at the appropriate difficulty level for each individual student to facilitate valid and reliable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

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

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



ACCESS for ELLs Online: Listening and Reading



ACCESS for ELLs Online: Speaking and Writing

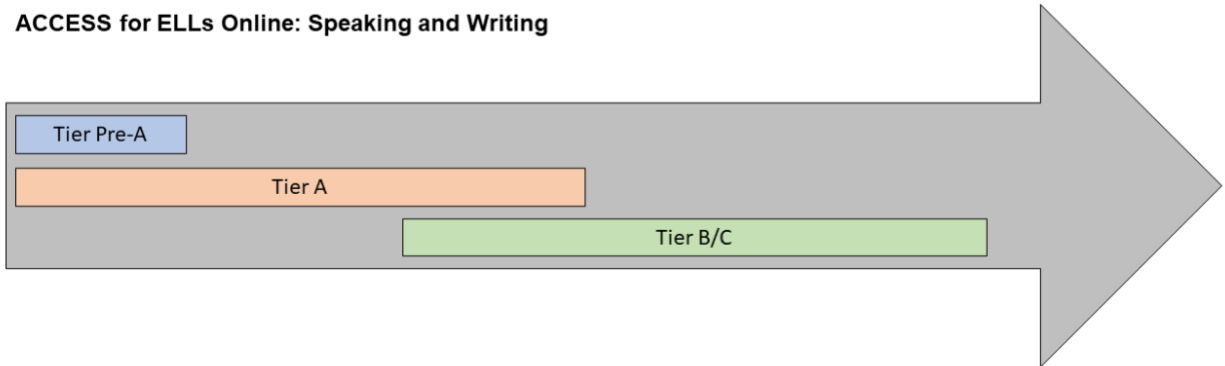


Figure 2. Tiers and proficiency levels.

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.

When the task models for ACCESS Online were first developed, CAL and WIDA accounted for issues of fairness by ensuring that principles of Universal Design of Assessments (UDA) were adhered to in this design phase (National Center on Educational Outcomes, 2021). 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 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). 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 for Reading items is 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, though some operational items and some items embedded for field testing purposes involve enhanced item presentations, including hot spot and drag-and-drop items (i.e., the student either clicks on an area of the screen or drags an image/text to a specified screen area to respond). 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.

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.

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.

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 3 shows the general screen layout of the Speaking test.

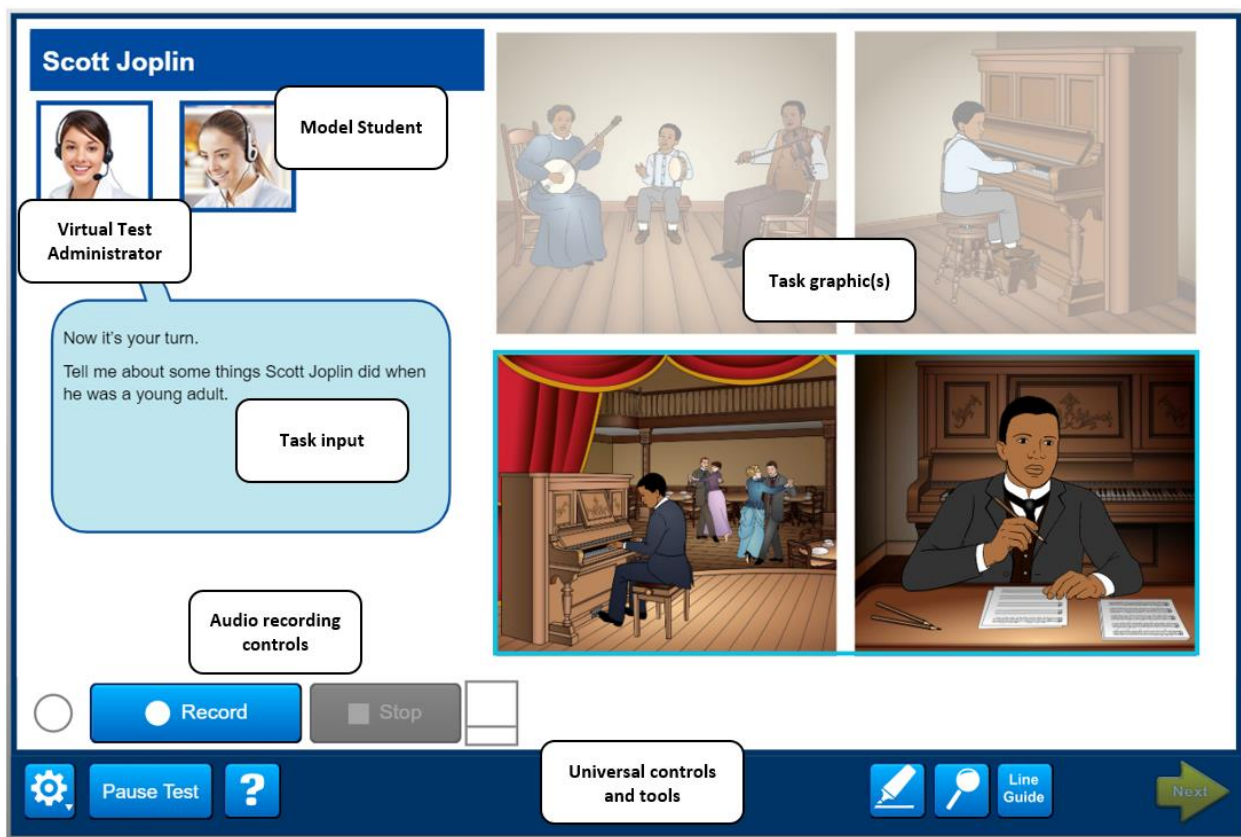


Figure 3. Visualization of the Speaking test screen layout.

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 Online is assembled to ensure that the evidence collected is (a) sufficient to make the required decisions based on the test results, and (b) appropriate for the student’s level of proficiency. 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 502 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	89%	0%	11%		
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	89%	0%	11%		
2–3	C	18	PL3–PL5	100%	0%	0%		
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	95%	5%	0%		
6–8	Entry	6	PL1–PL4	83%	0%	17%	Dichotomous selected response	Machine scored
6–8	A	12	PL1–PL3	92%	0%	8%		
6–8	B	18	PL2–PL4	55%	28%	17%		
6–8	C	18	PL3–PL5	89%	0%	11%		
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	84%	5%	11%		
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 4. 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 student’s ability is estimated based on performance on those six items, and that ability estimate is used 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 4); 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 4); 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 level of the next folder to be administered is chosen by the engine accordingly, 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 students whose ability estimate falls below PL 2.5 after the sixth stage end the test at this 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 4. 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 502 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	96%	4%	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	95%	0%	5%		
4–5	B	24	PL2–PL4	100%	0%	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%		
9–12	C	24	PL3–PL5	100%	0%	0%		

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

Figure 5 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. Students whose ability estimate falls below PL 2.5 after the eighth stage end the test at that point. Although timing guidance is included in the Test Administrator Manual, the Reading test is untimed.

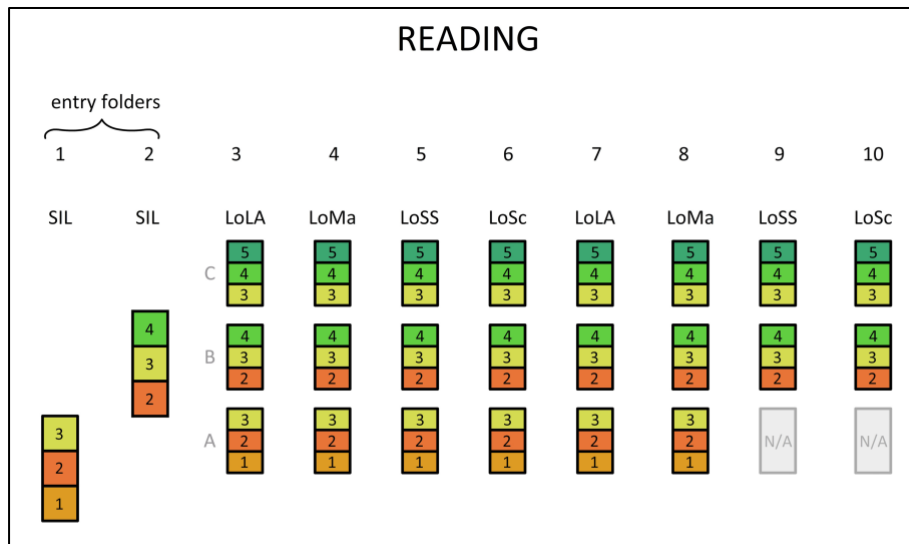


Figure 5. Format of the Reading test.

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	Human scored: centrally
6–8	B/C	2	PL2–PL5			

					response booklet or keyboarded in test platform	scored by DRC
9–12	A	2	PL1–PL3	Writing constructed response	Polytomous constructed response; handwritten in response booklet or keyboarded in test platform	Human scored: centrally scored by DRC
9–12	B/C	2	PL2–PL5			

As shown in Figure 6, 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.

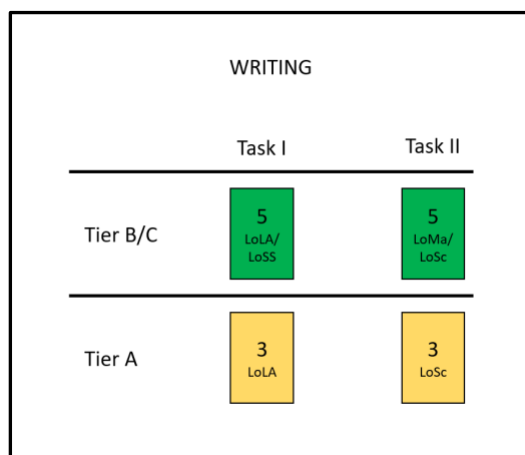


Figure 6. Format of the Writing test.

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 (Language of Language Arts and Language of Science, in that order), while 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.⁴ 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 502 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 are scored automatically by the test engine). 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 computer 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 There are two exceptions to the distribution of the WIDA Standards on the Series 502 Writing test. For Grade 1, Tier A, Task II is written to the Social and Instructional WIDA Standard. This task is a holdover from the transition to the two-task operational test design and was refreshed for Series 503. For Grades 6–8, Tier B/C, Task I is written to target Social and Instructional Language, Language of Language Arts, and Language of Social Studies. This item specification, previously called an Integrated Task, or IT task, was discontinued from development, but we were unable to refresh this slot in Series 501. We were able to refresh this slot in Series 503.

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 7 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 7, 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.

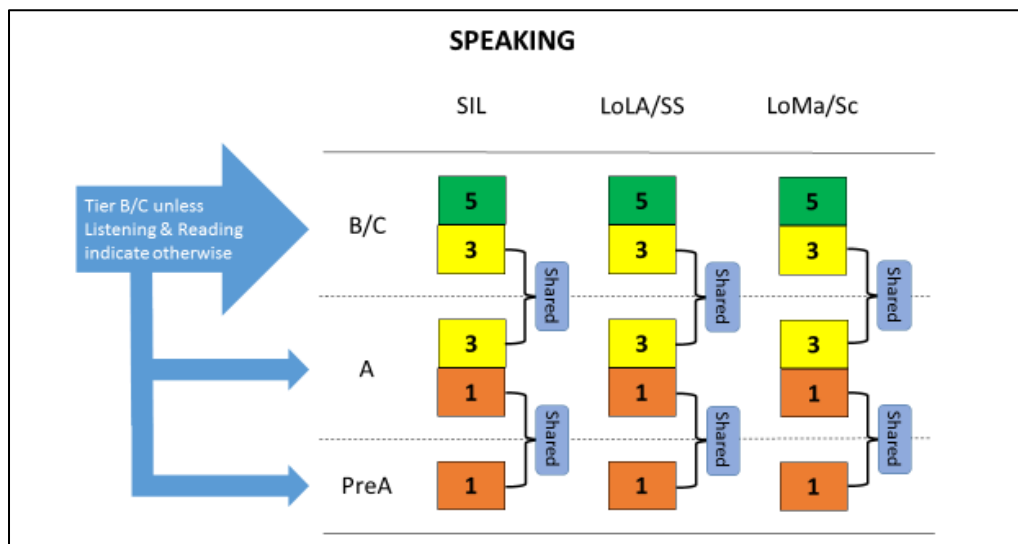


Figure 7. Format of the Speaking test.

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. CAL Psychometricians 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 to determine which tier of the Speaking test is administered 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 Development

The ACCESS item development process spans approximately 3 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 performance, length of time that folders have been on the test, item-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 WIDA for approval.

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

- The CAL TD team analyzes prior items that could not be used operationally due to fit issues, or in cases where the fit was acceptable, a difficulty 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 mechanics (e.g., an issue with the wording of the passage or stem, a distractor that is too attractive) or if a deeper item-specification issue is at fault (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 specifications as content standards change. As noted above, the ACCESS item specifications include explicit connections to the content standards. If an update to the relevant content standard makes an ACCESS item specification obsolete, the CAL TD team revises or replaces the specification.

Once updates to item specifications are complete, item development begins. The generation of initial item content occurs in two interconnected steps. First, the CAL TD team initiates a process of theme generation. In the ACCESS item 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 specification for Grades 4–5 is U.S. history, an example of an appropriate theme might be “the Industrial Revolution.”

CAL and WIDA recruit 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 CAL provides these educators with key parts of the item 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, that 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 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 CAL TD team then assign themes to professional item writers to develop the initial item content. CAL recruits individuals with prior experience developing ESL or English language arts items, preferably in the context of large-scale, standardized assessments, but individuals with other experience (such as experience writing items for language tests in languages other than English, and experience with English placement tests for the college/university setting) are also considered. All item writers, both new item writers and those returning from the previous test development cycle, participate in an introductory training, and CAL provides them with extensive documentation regarding writing items for ACCESS, including an Item Writing Handbook and ancillary documents (i.e., checklists, item specifications, templates) to complete

their assignments. One or more CAL Language Testing Specialists work with each item writer, providing feedback on the item content.

After item 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 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 writing. CAL and WIDA jointly recruit 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 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 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, 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 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, they are inserted into the folders, and layout review and fact-checking are conducted (with Test Development Manager sign-off) to ensure that the items 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 staff carry out on ACCESS items. WIDA assembles these panels by recruiting educators of multilingual learners from around the consortium, including culturally, racially, and linguistically diverse educators who reflect the population of students that take WIDA assessments. WIDA involves several criteria in the selection process which differ slightly between content review and bias and sensitivity review.

Content reviews occur by grade cluster (G1, G2-3, G4-5, G6-8, and G9-12) and the educators who are recruited to review a particular grade cluster's content (4 reviewers per grade cluster)

have experience teaching English language learners and are either currently teaching that grade cluster or have extensive experience teaching that grade cluster. Further criteria are used to try to ensure a good balance within the panels. These criteria include recruiting at least one educator within each panel with experience in each of the following areas: ELA, Science, Math, Social Studies, Special Education. Additionally, during the recruitment process, WIDA seeks to ensure diversity and balance across a) consortium states, b) locale (rural/suburban/urban), c) educator background, and d) years of teaching experience. CAL and WIDA first train the Content Review Panel on the procedures and scope of the review. The panelists are introduced to the test layout, are instructed on the logistics of the review, and are trained on the use of the review checklist. The panel then reviews each item and task to determine whether the content is accessible and relevant to students in the targeted grade-level cluster and at the targeted WIDA proficiency level, and that each item or task 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 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. Bias and sensitivity reviews occur by grade groupings (e.g., G1-3, G4-5, G6-8, and G9-12) and the educators who are recruited to review a particular grade cluster's content (5 or 6 reviewers per grade grouping) are educators or administrators who have experience teaching English language learners and are either currently teaching the grades within their group or have experience teaching those grades. Further criteria are used to ensure balance within the panels, as a variety of perspectives is crucial for the bias and sensitivity reviews. These criteria include recruiting at least one educator per panel with experience in Special Education. Additionally, during the recruitment process, WIDA seeks to ensure diversity and balance across a) consortium states, b) locale (rural/suburban/urban), c) educator background, and d) years of teaching experience. CAL and WIDA conduct training for all new and returning reviewers before any items are reviewed. CAL and WIDA staff facilitate the synchronous reviews and take extensive notes to capture all feedback during the reviews. WIDA TD staff also conducts a separate, asynchronous review around the time of the Content Review and Bias and Sensitivity Review, using the same materials that the educators review, and provides written feedback on the materials.

Once all Content Review and Bias and Sensitivity Review feedback from educators and from WIDA has been compiled, CAL Language Testing Specialists work to implement the feedback, with CAL Test Development Manager sign-off as a final step. Graphics and the QTI packages are subsequently revised by the CAL Test Production and Tech teams accordingly. The input and feedback from educators at various stages in the item development process serves as evidence that each item is appropriate for the age and grade-level cluster for which it is intended.

Tasks in the domain of Writing 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 item specifications to better align the Writing tasks with classroom practice, these tryouts allow CAL to evaluate whether the Writing tasks will effectively elicit language at the targeted WIDA proficiency levels. For the Writing tryouts, CAL and WIDA jointly recruit educators with appropriate numbers of students at the targeted proficiency levels (approximately 15 students per task) to participate. The educators administer the tasks to their students and send the written responses back to CAL for analysis. The students and the educators also fill out short surveys about the tasks. 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 CAL and WIDA conduct quality control checks of the QTI. WIDA signs 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 item development, the CAL TD team focuses on issues of fairness. First, the team applies UDA principles during item development. At the item specification level, the CAL TD team aims to precisely define the construct that each item or task is intended to measure. For the linguistic content of items, several principles of UDA are built into item development checklists and are specifically reviewed by CAL's TD managers and external reviewers (including WIDA TD staff and outside educators during Standards Expert Review and Bias and Sensitivity and Content Review), including

- Accessible, nonbiased items
- Amenability to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

In recent years, WIDA has placed additional focus on ensuring that the items, and especially the graphics, are amenable to accommodations by involving WIDA's Accessibility and Accommodations Team directly in the item 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 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 ensures that ACCESS Online items 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 502 as embedded folders during the operational administration of Series 501. 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 502, DRC field tested a total of 105 Listening items (35 folders), across all five grade-level clusters, as indicated in Table 5.

Table 5

Number of Field Test Folders and Items for the Series 502 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	1	2	6	SIL
1	A	1	0	1	3	LoLA
1	B	1	1	2	6	LoLA
1	C	1	1	2	6	LoLA
2-3	Entry	1	1	2	6	SIL
2-3	A	1	0	1	3	LoSS
2-3	B	1	1	2	6	LoSS
2-3	C	1	1	2	6	LoSS
4-5	Entry	1	1	2	6	SIL
4-5	A	1	0	1	3	LoLA
4-5	B	1	1	2	6	LoLA
4-5	C	1	1	2	6	LoLA
6-8	Entry	1	1	2	6	SIL
6-8	A	2	1	3	9	LoLA, LoSS
6-8	B	0	0	0	0	
6-8	C	1	1	2	6	LoSS
9-12	Entry	1	1	2	6	SIL
9-12	A	1	0	1	3	LoSS
9-12	B	0	0	0	0	
9-12	C	2	2	4	12	LoLA, LoSS
Total		20	15	35	105	

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 was students who were routed to the tier that the field test folder targeted, and the other 50% was 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 was 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 Psychometricians 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.

After CAL Psychometricians 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.

Each item was classified according to 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 discontinued from consideration. If one or more items were yellow, the Post–Field Test Review Panel reviewed the content of each item, along with relevant statistics like distractor analyses, to determine if the items 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 502

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

2.3.2.2. Reading

DRC field tested the Reading items developed for Series 502 as embedded items during the operational administration of Series 501. 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 502, DRC field tested a total of 186 Reading items (62 folders), across all five grade-level clusters, as indicated in Table 7.

Table 7

Number of Field Test Folders and Items for the Series 502 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	1	2	6	SIL
1	A	4	2	6	18	LoLA, LoMa
1	B	1	1	2	6	LoMa
1	C	0	0	0	0	
2–3	Entry	1	1	2	6	SIL
2–3	A	0	0	0	0	
2–3	B	2	1	3	9	LoLA, LoSS
2–3	C	3	2	5	15	LoLA, LoSS
4–5	Entry	1	1	2	6	SIL
4–5	A	3	1	4	12	LoLA, LoMa, LoSc
4–5	B	0	0	0	0	
4–5	C	3	3	6	18	LoLA, LoSS, LoSc
6–8	Entry	1	1	2	6	SIL
6–8	A	3	2	5	15	LoLA, LoMa, LoSS
6–8	B	2	2	4	12	LoSS, LoSc
6–8	C	2	2	4	12	LoSS, LoSc
9–12	Entry	1	1	2	6	SIL
9–12	A	3	2	5	15	LoLA, LoMa
9–12	B	2	2	4	12	LoLA, LoMa
9–12	C	2	2	4	12	LoLA, LoMa
Total		35	27	62	186	

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 Psychometricians 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 502 Writing tasks in an embedded field-test model for the first time. For Series 502, a total of 13 Writing tasks were field tested, as indicated in Table 8.

Table 8
Number of Field Test Tasks for Series 502 Writing

Grade-Level Cluster	Tier	Number of Folders to Refresh	Number of Folders Field Tested	Standards Addressed in FT
1	A	1	1	LoSc
1	BC	1	1	LoMa/LoSc
2–3	A	1	1	LoSc
2–3	BC	1	1	LoMa/LoSc
4–5	A	1	1	LoSc
4–5	BC	1	2	LoMa/LoSc
6–8	A	1	1	LoSc
6–8	BC	1	2	LoMa/LoSc
9–12	A	1	1	LoSc
9–12	BC	1	2	LoMa/LoSc
Total		10	13	

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 nine-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.⁵

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 54 tasks (18 panels) for Series 502, with a target sample size of 500 students per folder. This is well more than the minimum of 250 per form for high-stakes tests that Linacre (1994) proposed, and allows for at least 10 observations per category, as recommended by Linacre (2002a) for polytomous items. Since the score distribution for Speaking is highly concentrated in the middle of the distribution, with relatively fewer percentage of cases at the high end of the distribution, a sample size of 500 was chosen to ensure that there will be students at the high end of the score distribution for analysis, as well as to ensure that students' Speaking performances are available at those score points to create scoring materials.

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 502, CAL field tested a total of 36 Speaking tasks, as indicated in Table 9.

⁵ 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 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.

Table 9

Number of Field Test Tasks for Series 502 Speaking

Grade-Level Cluster	Tier	Number of Folders to Refresh	Number of Folders Field Tested	Standards Addressed in FT
1	A	2	4	SIL, LoMa/LoSc
1	BC	2	4	SIL, LoMa/LoSc
2–3	A	1	2	SIL
2–3	BC	1	2	SIL
4–5	A	2	4	SIL, LoMa/LoSc
4–5	BC	2	4	SIL, LoMa/LoSc
6–8	A	2	4	SIL, LoMa/LoSc
6–8	BC	2	4	SIL, LoMa/LoSc
9–12	A	2	4	SIL, LoMa/LoSc
9–12	BC	2	4	SIL, LoMa/LoSc
Total		18	36	

2.3.3. Item Review and Selection

After the analysis of field test data, a panel consisting of WIDA and CAL staff conducted an item selection meeting to determine which of the field-tested folders would be placed on the Series 502 operational assessment. Results from qualitative and quantitative analyses guided the selection of operational items.

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), and CAL 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 and infit 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 distractor analyses, 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.

Tables 10 and 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 502 Listening Test, by Grade-Level Cluster

Grade-Level Cluster	Number of New Items	Number of Continuing Items	Total Number of Items
1	9	45	54
2–3	3	51	54
4–5	6	48	54
6–8	12	42	54
9–12	18	36	54

Table 11

Number of New and Continuing Items on ACCESS Online, Series 502 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	15	57	72
4–5	21	51	72
6–8	15	57	72
9–12	15	57	72

In the domains of Writing and Speaking, the Item 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 field-tested tasks. The CAL TD team then integrated the observations with task statistics, including fit statistics, raw score distributions, and rater agreement indices, to produce a recommendation for the panel regarding (a) in cases where there was a choice between two tasks, which task to place on the operational test, or whether to leave that slot unrefreshed; or (b) in cases where there was only a single task, whether to place the task on the operational test or to leave that slot unrefreshed.

Although rater agreement indices and fit statistics were considered in the recommendations, the CAL TD team based their recommendations primarily on the qualitative data (i.e., whether students could successfully score in the intended range and whether DRC raters observed major anomalies that could indicate that the tasks were not performing as intended), the raw score distributions and the difficulty measures, as the field-test tasks and the operational tasks up for refreshment should have comparable statistics in these two areas. 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 difficulty measures and the raw score distributions of the field-test tasks varied too much from those of the operational tasks.

Tables 12 and 13 provide the numbers of continuing and new items, per grade-level cluster, for the Writing and Speaking 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 12

Number of New and Continuing Items on ACCESS Online Series 502 Writing Test, by Grade-Level Cluster

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

Table 13

Number of New and Continuing Tasks on ACCESS Online Series 502 Speaking Test, by Grade-Level Cluster

Grade-Level Cluster	Tier	Number of New Items	Number of Continuing Items	Total Number of Items
1	Pre-A	2	1	3
1	A	4	2	6
1	B/C	4	2	6
2-3	Pre-A	2	1	3
2-3	A	4	2	6
2-3	B/C	4	2	6
4-5	Pre-A	2	1	3
4-5	A	4	2	6
4-5	B/C	4	2	6
6-8	Pre-A	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. During the 2020 school year, many states extended their testing windows due to the COVID-19 pandemic. 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 or record their answers. Once a student records 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.

3.1.2. Writing

Students in Grades 1–3 perform the Writing tasks on paper. All students in Grades 1–3 handwrite a 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 your 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 training course within the WIDA Secure Portal (<https://grow.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 your state's testing window. Training courses include test preparation and administration tutorials and online administration quiz.

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

For detailed guidelines about training and test administration, please refer to the *ACCESS for ELLs Test administrator manual* and the *ACCESS for ELLs District and School Test Coordinator Manual*.

3.2.1 Administering the Test Practice

The administration of the practice test for an individual test domain 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 practice test again if needed. The narration within the practice test 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.

3.2.2 Listening Test Administration

The Listening test (including test practice items) is designed to take approximately 30 to 40 minutes. Note that the approximate test administration time does not include convening students, taking attendance, or explaining test directions.

3.2.3 Reading Test Administration

The Reading test (including directions and practice items) is designed to take approximately 35 minutes. Note that the approximate test administration time does not include convening students, taking attendance, or explaining test directions.

3.2.3.1 Reading Test Item Types

The Reading test may include three different item types: multiple choice, hotspot, and drag and drop. Although a student may not see all three of these item types, it is important to ensure that students know what to do for these different item types.

- **Multiple choice.** Students choose an answer from a set of ordered response options under the question. 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.
- **Hotspot.** Students see a large response area under the question. 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.
- **Drag and drop.** There are two examples of this item type. 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.

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.

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. Once the students have completed the Listening and Reading tests, Test Coordinators run a Tier Placement Report that identifies the tier each student is assigned to take. Test administrators use the report to know which form to administer to which student. 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.

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. The tier the student takes is determined by the student's Listening and Reading test results and automatically loads for the student upon logging into the test platform with test ticket information. 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 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 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.

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 individual accessibility needs. These may either be embedded in the online test or provided by Test Administrators during testing. Universal tools do not affect the construct being measured on the assessment.

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.

Accessibility features include tools that are available to all ELLs taking ACCESS for ELLs. Examples of accessibility features include highlighter, line guide, magnification, and color overlay. All accessibility features are available to all ELLs during testing; specific designation

is not required prior to testing to make them available to the student during 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.

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.

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 test taker 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 score 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 the performance-based tasks 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 holds recruiting events, after which applications for rater positions are screened by DRC’s recruiting staff and likely candidates are personally interviewed by DRC staff. As part of the hiring process, DRC requires 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 gives preference to candidates with previous experience scoring large-scale assessments and degrees emphasizing expertise in English language arts. The rater pool consisted of educators, writers, editors, and other professionals with content-specific backgrounds. While DRC valued these individuals for their content-specific knowledge, they were required to set aside their own biases about student performance and accept the scoring standards outlined in the training for scoring the ACCESS for ELLs.

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

Rater Training and Qualifying

- DRC assigns each rater a unique ID number and password.
- The scoring director provides detailed directions for use of DRC’s computerized scoring system and remote communication tools.

- The scoring director trains the raters using task-specific anchor sets and training sets.
- Raters must demonstrate scoring proficiency by scoring at least 70% agreement on a qualifying set before scoring live responses.
- Once raters are qualified, DRC provides further training for their grade-level cluster and on the specific tasks for which they will rate responses.
- Once raters have trained, qualified, and begun live scoring, DRC uses calibration sets (of which there are two types, recalibration sets and validation sets, which we explain below) to keep the raters calibrated on the actual tasks they are scoring.

Calculating Score Agreement for Score Monitoring

- DRC’s handscoring system generates handscoring reports, detailing agreement rates for each rater and item. These reports are customized based on input and direction from WIDA. The reports are automatically generated overnight throughout the course of handscoring and may also be run on demand. DRC provides weekly interrater reliability reports to WIDA throughout the handscoring process to ensure that DRC maintains sufficient quality control throughout the course of scoring.
- For Writing, we define **agreement** as two adjacent scores, reported as %AG. (See Section 3.2.3 for a description of the Writing Scoring Scale.) For example, using the Writing Scoring Scale, we consider scores of 2 and 2+ as agreement, as well as scores of 2 and 2 or scores of 2+ and 3. However, we consider scores of 2 and 3 on the Writing Scoring Scale as **adjacent**, while we consider scores of 2 and 3+ as **nonadjacent**.
- For Speaking, we define **agreement** as two scores that are exactly the same, reported as %EX. (See Section 3.2.4 for a description of the Speaking Scoring Scale.) Unlike in Writing, where DRC considers two adjacent scores as “Agreement,” Speaking raters 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 Ratings

- The DRC scoring system routes and reroutes responses to raters until enough raters perform the prescribed number of ratings for all responses.
- Raters do not see the scores of the other raters and do not know if they are the first or second rater.
- The purpose of the first and second ratings is to monitor interrater reliability by comparing the scores given by two separate raters to the same response. When calculating final scores, the first score given is the score of record.

Monitoring Scoring (Quality Control)

Ongoing quality control checks and procedures help monitor and maintain the quality of the scoring sessions. DRC’s handscoring reports are automatically generated overnight and are also

available on demand to monitor progress and maintain handscoring quality control. DRC provides WIDA with access to these reports on a regular basis throughout the scoring process to provide assurance that the quality control metrics meet or exceed expectations.

- During the handscoring process, the scoring directors communicate 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 conduct routine read- and listen-behinds to observe, in real time, raters' performance. Team leaders utilize live, scored responses to provide ongoing feedback and, if necessary, retraining for raters.
- The scoring system randomly selects at least 20% of tasks for two raters to independently score, for the purpose of monitoring interrater reliability. Raters are not aware that another rater may have previously scored a task.
- The DRC system generates interrater reliability reports daily to monitor how often each rater's scores match other raters' scores, and scoring leaders continually monitor individual statistics compared to the group average. If the agreement rate for a rater falls below 70%, supervisors increase monitoring and retraining activities with the rater. If the rater fails to demonstrate improved reliability, the rater is released from scoring the item.
- Since the interrater agreement rates were all at or above 70%, the target stipulated by WIDA, the focus turned to raters with lower-than-average agreement rates—even if their agreement was at or above 70%. Even when all agreement rates are at or above 70%, scoring supervisors continue to seek opportunities to increase reliability by providing ongoing feedback and retraining to the raters based on the specific performance of each rater as evidenced by the quality control reports and observations made when reviewing scores given by raters to tasks.
- Responses can be retrieved on demand (e.g., specific grade-level clusters, specific students) should the need arise during or after the scoring process.
- If needed, responses can be rescored based on task- or response-level information, such as task number, date, score value 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 developed these recalibration sets and validity responses together. CAL developed an initial pool of responses for use as recalibration and validity 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 raters score a task, they take one recalibration set of five responses. The recalibration sets did not differ from rater to rater. For example, a recalibration set was specified for the first day that a rater scored a specific task; every rater who scored that task took this same recalibration set on the first day that they scored

that task. After the raters took the recalibration sets, 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 scoring a task, DRC used validity responses to continue monitoring rater performance. DRC seeded the validity responses into operational scoring so that the raters did not know which responses were operational and which were validity responses. Reports generated daily compared the scores given by each rater to the "true" score for each validity response. When a rater was working on a task, DRC dealt the validity responses to that rater in a random order. Each validity response was dealt to multiple raters over the course of the project (i.e., given enough time, every rater working on a task would score every validity response for that task), but the validity responses were not dealt in the same order to each rater.

Handling Unusual Responses

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

- **Scoring questions.** If raters 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 raters forwarded the response to team leaders for assistance. The team leaders then reviewed the response and applied the proper score. If anything about the response and the rater's question indicated that the rater needed any clarifications about the scoring guidelines, the team leaders met with raters to review the response and to explain how to score it based on the 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 constitutes a nonscorable response to ensure consistency of nonscorable codes, and this information was provided from CAL to DRC along with other item-specific training materials that were used to train DRC's raters. During scoring, when raters apply a nonscorable code (except for Blank), the response was automatically forwarded to a handscoring supervisor for review and approval. If the handscoring supervisors had any questions about the application of nonscore codes to specific responses, DRC contacted WIDA and CAL representatives for further review and discussion.
- **Alerts.** To handle possible alert papers (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 response was flagged with the alert status, it was automatically routed to handscoring supervisors for review. When the handscoring supervisors concurred with the "alert" status of the

response, the response was then passed on 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 would flag the response with a “request original” status. When a response was flagged as “request original,” it was automatically forwarded 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 request was forwarded to staff in DRC’s Operations Services, who located the original student response so that it could be reviewed by handscoring supervisors to score the response.

Remote Scoring Procedures due to the COVID-19 Pandemic

Prior to 2020, all WIDA handscoring was conducted in DRC’s handscoring centers. 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 done in the physical scoring locations. The platform, content, and expectations for quality remained the same, and interactive technology and content training and discussions were conducted live (virtually). The differences came with the method through which DRC delivered training (online) and in the modes of communication used (web screen sharing, webcast, video chat, and chat). DRC equipped scoring leaders with a variety of tools to ensure 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.

Live, interactive training was conducted via Moodle Learning Management System, which mirrors aspects of the scoring room and provides 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). Printing was disabled for raters in ScoreBoard to protect the security of the student responses, test questions, and training materials. Restrictions built into ScoreBoard defined the hours during the day 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.
- **Content training with Moodle.** DRC provided content training remotely as an interactive, comprehensive, hands-on experience. For Writing training, scoring directors trained groups of raters by screensharing PDFs of training materials. Each training example was viewed individually, with supervisors directing scorers 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 be redealt and rescored.

4.3. Writing Scoring Scale

The Writing Scoring Scale has six whole score points that range from 1 to 6. For responses that fall in between the whole score points, “plus” score points are available (e.g., a response that falls between 3 and 4 is scored as 3+). 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 8 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
	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
	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
3+	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
	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
<p style="text-align: center;">D: Discourse Level S: Sentence Level W: Word/Phrase Level</p>		

Figure 8. 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, a whole score point is awarded. If there is clear evidence that one or two descriptors from an adjacent score point are a better fit, the rater awards a plus score point between the two applicable whole score points.

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, as defined below.

Nonscorable: The response is blank; consists only of verbatim copied text; consists only of text that is completely off task; is entirely in a language other than English; or appears to have been plagiarized from an outside source during testing.

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. These responses are scored in their entirety using the scoring scale; however, the maximum holistic score for a completely off-topic response is 2+.

Partially off-task response: The response contains both off-task and on-task writing. These responses are scored 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). These responses are scored in their entirety using the scoring scale.

Both nonscorable and completely off-task responses are scored as 0. Completely off-topic responses receive a maximum score of 2+. Partially off-topic responses are scored in their entirety, while partially off-task responses are scored by ignoring the off-task portion of the response and scoring only the on-task portion.

To calculate a raw score for the Writing test, raters' scores for each Writing task are converted to whole numbers ranging from 0 to 9, as shown in Table 14. Raw scores for the two operational tasks are added, giving 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. The Writing Scoring Scale was designed specifically as a scoring tool and 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. 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.).

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

English language use that is equal to or beyond the English language use illustrated by the model student’s response.

Figure 9 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	<ul style="list-style-type: none"> • Language use comparable to or going beyond the model in sophistication • Clear, automatic, and fluent delivery • Precise and appropriate word choice
Strong use of oral language to provide a detailed response	<ul style="list-style-type: none"> • Language use approaching that of model in sophistication, though not as rich • Clear delivery • Appropriate word choice
Adequate use of oral language to provide a satisfactory response	<ul style="list-style-type: none"> • Language use not as sophisticated as that of model • Generally comprehensible use of oral language • Adequate word choice
Attempted use of oral language to provide a response in English	<ul style="list-style-type: none"> • Language use does not support an adequate response • Comprehensibility may be compromised • Word choice may not be fully adequate
No response (in English)	<ul style="list-style-type: none"> • Does not respond (in English)

Figure 9. Speaking Scoring Scale.

The Speaking Scoring Scale includes descriptors for overall language use, response sophistication, language delivery, and word choice. As stated above, the scale is applied relative to the proficiency level demands of the task. For tasks targeting language elicitation at PL 1, there are only three possible score points: *No Response*, *Attempted*, and *Adequate and Above*. This is the case because appropriate responses to PL 1 tasks are single words and short chunks of language, so it is not possible to reliably distinguish between *Adequate*, *Strong*, and *Exemplary* performances.

To calculate a raw score for the Speaking test, the five score points are converted to whole numbers, as shown in Table 15. To calculate a total raw score, the raw scores for each task are added together; additionally, in Tier B/C, six points are added to the total raw score, representing a score of *Adequate and Above* for three tasks targeting language at PL 1. Though a Tier B/C student would not be administered any tasks targeting the PL 1 level, it is assumed that a student who had been routed to the B/C test would easily achieve a score of *Adequate and Above* on these tasks. Thus, on the Pre-A test, scores can range from 0 to 6; on the A test, from 0 to 18; and on the B/C test, from 6 to 30.

Table 15

Rating to Raw Score Conversion (Speaking)

Rating	Raw score
No Response (R, F, or I)*	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.

Speaking tasks are scored 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. The Speaking Scoring Scale was designed specifically for test scoring use and is not intended for classroom purposes.

5. Summary of Score Reports

5.1. Individual Student Report

Score reports (district, school, and student level reports) are made available in the WIDA Assessment Management System (AMS) as soon as they are available for each state and printed reports are shipped to school districts and schools at the same time or shortly thereafter. Score reports are available for states to identify students' language performance and properly determine language support for ELLs. Each state and school district determines when and how students individual score reports are provide to students' parents or guardians. Communication about student score reports and resources that districts use to support interpretation is a local decision. WIDA provides resources that schools, districts and states may use to aid in score interpretation. (See below.) How that material is used 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-0050568bee8>).

According to Kim et al., (2016; 2020), educators find interpreting technical terms to be challenging, which suggests the need for describing terms with more clarity in score reports. WIDA plans to evaluate current score reports through focus groups to identify how

improvements can be made to help educators, families, and students to better understand score information.

The Individual Student Report (Figure 10) contains detailed information about the performance of a single student within Grades K–12. Its primary users are students, parents/guardians, teachers, and school teams. It describes the language needed to access content and succeed in school, one indicator of a student’s English language proficiency.







ACCESS for ELLs 2.0*
English Language Proficiency Test

Sample Student
Birth Date: mm/dd/yyyy | Grade: sample grade
Tier: sample tier
District ID: XXXXXXXXXXXXXXXX | State ID: XXXXXXXXXXXXXXXX
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 (Possible 1.0-6.0)						Scale Score (Possible 100-600) and Confidence Band See Interpretive Guide for Score Reports for definitions					
	1	2	3	4	5	6	100	200	300	400	500	600
Listening 	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: <ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> • 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 10. Individual Student Report.

The score report includes four domain scores (Listening, Speaking, Reading, and Writing) and four composite scores (Oral Language, Literacy, Comprehension, and Overall). Each composite score is represented by a label, a breakdown of how individual domains are used to calculate it, and a visual display of the results. Composition of single domain scores in composite scores is presented in the individual student report. 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 requests (Sahakyan, N., 2020).

The proficiency level is presented both graphically and as a whole number followed by a decimal. The shaded bar of the graph reflects the exact position of the student's performance on the 6-point English Language Proficiency Scale. The whole number reflects a student's English language proficiency level (1–Entering, 2–Emerging, 3–Developing, 4–Expanding, 5–Bridging, and 6–Reaching) in accord 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 associated confidence band. The confidence band reflects the standard error of measurement of 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 is equal to the 95% probability level.

If a student does not complete one or more of the language domains, NA (not available) is inserted in that language domain as well as in all applicable composite scores, including the overall score. Students with identical overall scores may have very different profiles in terms of their Listening, Speaking, Reading, and Writing.

The second part of the Student Report provides information about the individual student's proficiency levels as whole numbers and describes what students at the reported proficiency level may typically be expected to be able to do in English. For example, if the student received a proficiency level score of 2 for Speaking, the report will include a description of the type of spoken language the student may be expected to be able to produce.

When interpreting scores, the following points should be kept in mind by all stakeholders:

- 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 the individual test.
- No single 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 from different domains should not be compared. 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 level scores can be used to compare test scores 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, identical to those 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 attain each proficiency level within a given population.

Part 2:
Technical Results

Contents

1	Annual Test Results	1-5
1.1	Participation	1-6
1.1.1	Grade-Level Cluster	1-6
1.1.2	Grade	1-9
1.2	Scale Score Results	1-14
1.2.1	Mean Scale Score Across Domain and Composite Score by Cluster	1-14
1.2.2	Mean Scale Score Across Domain and Composite Score by Grade	1-19
1.2.3	Correlations	1-28
1.3	Proficiency Level Results	1-30
1.3.1	Domains	1-30
1.3.2	Composites	1-38
2	Analysis of Domains	2-1
2.1	Complete Item or Task Analysis and Summary	2-4
2.1.1	Listening	2-7
2.1.2	Reading	2-17
2.1.3	Writing	2-32
2.1.4	Speaking	2-42
2.2	DIF Analysis and Summary	2-47
2.2.1	Listening	2-50
2.2.2	Reading	2-52
2.2.3	Writing	2-55
2.2.4	Speaking	2-58
2.3	Raw Score Distribution for Speaking and Writing	2-62
2.3.1	Listening	2-62
2.3.2	Reading	2-63
2.3.3	Writing	2-63
2.3.4	Speaking	2-68
2.4	Scale Score Distribution	2-78
2.4.1	Listening	2-79
2.4.2	Reading	2-84
2.4.3	Writing	2-89
2.4.4	Speaking	2-99
2.5	Proficiency Level Distributions	2-109
2.5.1	Listening	2-110
2.5.2	Reading	2-115

2.5.3	Writing.....	2-120
2.5.4	Speaking.....	2-135
2.6	Raw Score to Scale Score to Proficiency Level Conversion for Speaking and Writing 2-155	
2.6.1	Listening.....	2-155
2.6.2	Reading.....	2-155
2.6.3	Writing.....	2-156
2.6.4	Speaking.....	2-166
2.7	Equating Summary.....	2-176
2.7.1	Listening.....	2-182
2.7.2	Reading.....	2-192
2.7.3	Writing.....	2-205
2.7.4	Speaking.....	2-215
2.8	Test Characteristic Curve.....	2-220
2.8.1	Listening.....	2-221
2.8.2	Reading.....	2-221
2.8.3	Writing.....	2-222
2.8.4	Speaking.....	2-229
2.9	Test Information Function.....	2-240
2.9.1	Listening.....	2-243
2.9.2	Reading.....	2-245
2.9.3	Writing.....	2-248
2.9.4	Speaking.....	2-255
3	Analyses of Composite Scores.....	3-1
3.1	Scale Score Distribution for Composites.....	3-1
3.1.1	Oral.....	3-2
3.1.2	Literacy.....	3-7
3.1.3	Comprehension.....	3-12
3.1.4	Overall.....	3-17
3.2	Proficiency Level Distribution for Composites.....	3-22
3.2.1	Oral.....	3-23
3.2.2	Literacy.....	3-28
3.2.3	Comprehension.....	3-33
3.2.4	Overall.....	3-38
4	Annual Updates of Validity Evidence.....	4-1
4.1	Standards.....	4-2
4.1.1	Test Content.....	4-2

4.1.2	Response Processes	4-2
4.1.3	Internal Structure	4-2
4.1.4	Relation to Other Variables	4-2
4.2	Annual Validity Studies.....	4-3
4.2.1	Enhancement of ACCESS Online Tests in Comparison with ACCESS Paper Tests.....	4-3
4.2.2	Dimensionality and Content Knowledge in ACCESS Tests	4-3
5	Reliability.....	5-1
5.1	Reliabilities of the Domain Scores	5-6
5.1.1	Listening	5-10
5.1.2	Reading.....	5-11
5.1.3	Writing.....	5-12
5.1.4	Speaking	5-14
5.2	Interrater Agreement Rates.....	5-16
5.2.1	Listening	5-17
5.2.2	Reading.....	5-17
5.2.3	Writing.....	5-17
5.2.4	Speaking	5-20
5.3	Conditional Standard Errors of Measurement of the Scale Scores at the Cut Points 5-25	
5.3.1	Listening	5-27
5.3.2	Reading.....	5-30
5.3.3	Writing.....	5-33
5.3.4	Speaking	5-36
5.4	Accuracy and Consistency of Domains.....	5-39
5.4.1	Listening	5-45
5.4.2	Reading.....	5-46
5.4.3	Writing.....	5-48
5.4.4	Speaking	5-49
5.5	Reliabilities of Students' Composite Scores.....	5-51
5.5.1	Oral	5-54
5.5.2	Literacy.....	5-56
5.5.3	Comprehension.....	5-58
5.5.4	Overall	5-60
5.6	Conditional Standard Errors of Measurement for the Students' Composite Scale Scores 5-64	
5.6.1	Oral	5-66

5.6.2	Literacy	5-69
5.6.3	Comprehension	5-72
5.6.4	Overall	5-75
5.7	Accuracy and Consistency of Composites	5-78
5.7.1	Oral	5-82
5.7.2	Literacy	5-83
5.7.3	Comprehension	5-85
5.7.4	Overall	5-86
6	Quality Control	6-1
6.1	Content Development Quality Control	6-1
6.2	Test Administration Quality Control	6-3
6.3	Rater Quality Control	6-5
6.4	Score Reporting Quality Control	6-6
6.5	Data Forensic Quality Control	6-7

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 502 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, 2% to 7% of students were excluded in each cluster and domain.

In addition to these data exclusions, 161 student records were removed from the data set due to a concern over plagiarized responses on 9–12 Speaking and/or Writing tests. Further detail on this issue can be found in Section 6.5, Data Forensic Quality Control.

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	8,980	137,292	6.54%
	2–3	16,613	261,438	6.35%
	4–5	15,922	241,948	6.58%
	6–8	18,050	254,782	7.08%
	9–12	13,114	209,283	6.27%
	Total	72,679	1,104,743	6.58%
Reading	1	6,077	137,292	4.43%
	2–3	14,309	261,438	5.47%
	4–5	16,558	241,948	6.84%
	6–8	18,271	254,782	7.17%
	9–12	15,224	209,283	7.27%
	Total	70,439	1,104,743	6.38%
Writing	1	-	137,292	n/a
	2–3	-	261,438	n/a
	4–5	7,642	241,948	3.16%
	6–8	9,916	254,782	3.89%
	9–12	6,865	209,283	3.28%
	Total	24,423	1,104,743	2.21%
Speaking	1	8,764	137,292	6.38%
	2–3	17,611	261,438	6.74%
	4–5	16,232	241,948	6.71%
	6–8	18,759	254,782	7.36%
	9–12	13,843	209,283	6.61%
	Total	75,209	1,104,743	6.81%

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 38 WIDA states and U.S. territories that participated in the ACCESS Online operational testing program in 2020–2021 by grade-level cluster. The 38 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 38 states and U.S. territories. The state with the largest number of students was Georgia. The state/territory with the smallest number of

participants was the District of Columbia. The biggest cluster was Grade 1. The abbreviations are as follows: DC, District of Columbia; MP, Northern Mariana Islands; and BI, Bureau of Indian Education.

Table 1.1.1.1
Participation by Cluster by State, S502 Online

State	Cluster					Total
	1	2-3	4-5	6-8	9-12	
AK	493	963	1,062	1,226	1,033	4,777
AL	3,373	7,323	6,608	7,139	4,551	28,994
BI	27	82	97	134	6	346
CO	7,847	15,722	12,558	12,915	10,655	59,697
DC	10	11	10	7	1	39
DD	743	1,552	1,275	1,149	549	5,268
DE	1,354	2,759	2,378	2,212	1,324	10,027
GA	12,301	25,010	23,615	20,141	14,658	95,725
HI	1,547	3,478	3,150	3,251	2,409	13,835
ID	1,893	4,097	3,319	3,684	3,073	16,066
IL	14,169	13,475	25,431	24,716	17,172	94,963
IN	7,695	15,507	13,893	14,533	11,981	63,609
KY	3,781	6,385	4,901	4,529	4,532	24,128
MA	9,029	16,338	11,170	11,796	11,544	59,877
MD	666	1,548	1,758	1,528	805	6,305
ME	248	521	391	502	484	2,146
MI	6,376	13,694	12,083	13,636	14,465	60,254
MN	5,866	11,798	9,107	7,916	6,977	41,664
MO	3,760	7,376	5,851	5,484	5,062	27,533
MP	47	196	287	5	34	569
MT	247	484	513	704	370	2,318
NC	10,482	22,898	21,070	22,390	14,992	91,832
ND	383	739	660	725	748	3,255
NH	407	871	707	745	728	3,458
NJ	7,635	14,278	10,767	9,567	10,189	52,436
NM	212	623	1,026	1,138	531	3,530
NV	2,254	4,466	8,004	7,124	6,706	28,554
OK	6,126	12,677	11,415	11,856	8,634	50,708
PA	3,117	6,617	6,211	8,458	8,058	32,461
RI	1,318	2,489	2,392	2,935	3,225	12,359
SC	3,467	7,493	7,464	9,198	10,152	37,774
SD	723	1,339	1,157	1,266	912	5,397
TN	6,031	11,296	8,582	8,408	8,469	42,786
UT	4,317	9,981	10,396	12,482	8,062	45,238
VA	4,917	7,756	4,164	11,974	9,328	38,139
VT	149	309	299	280	311	1,348
WI	4,044	8,739	7,712	8,585	6,082	35,162
WY	238	548	465	444	471	2,166
Total	137,292	261,438	241,948	254,782	209,283	1,104,743

Table 1.1.1.2 shows participation by grade-level cluster by gender across all 38 states and U.S. territories combined, while Table 1.1.1.3 shows participation by grade-level cluster by ethnicity across all 38 states and U.S. territories. The gender ratio was generally 46% female and 52% male in Clusters 1–3 and 44% female and 54% male for Clusters 4–12. About 65% of participants were Hispanic in all clusters.

Table 1.1.1.2

Participation by Cluster by Gender, S502 Online

Cluster		Gender			Total
		F	M	Missing	
1	Count	63,966	71,773	1,553	137,292
	% within Cluster	46.59%	52.28%	1.13%	100.0%
2–3	Count	121,123	137,686	2,629	261,438
	% within Cluster	46.33%	52.66%	1.01%	100.0%
4–5	Count	108,655	130,714	2,579	241,948
	% within Cluster	44.91%	54.03%	1.07%	100.0%
6–8	Count	109,308	142,062	3,412	254,782
	% within Cluster	42.90%	55.76%	1.34%	100.0%
9–12	Count	90,912	115,475	2,896	209,283
	% within Cluster	43.44%	55.18%	1.38%	100.0%
Total	Count	493,964	597,710	13,069	1,104,743
	% within Cluster	44.71%	54.10%	1.2%	100.0%

Table 1.1.1.3

Participation by Cluster by Ethnicity, S502 Online

Cluster		Ethnicity			Total
		Hispanic	Non-Hispanic	Unknown	
1	Count	86,941	42,148	8,203	137,292
	% within Cluster	63.33%	30.70%	5.97%	100.0%
2–3	Count	169,023	76,528	15,887	261,438
	% within Cluster	64.65%	29.27%	6.08%	100.0%
4–5	Count	161,892	61,314	18,742	241,948
	% within Cluster	66.91%	25.34%	7.75%	100.0%
6–8	Count	173,071	58,969	22,742	254,782
	% within Cluster	67.93%	23.14%	8.93%	100.0%
9–12	Count	135,949	53,844	19,490	209,283
	% within Cluster	64.96%	25.73%	9.31%	100.0%
Total	Count	726,876	292,803	85,064	1,104,743
	% within Cluster	65.8%	26.5%	7.7%	100.0%

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 Cluster 2–3 percentages of Tier A became smaller. In the Speaking domain, percentages of Tier A remained smaller than Tier B/C for all clusters. Percentages of Pre-A in Speaking were 1% to 6%.

Table 1.1.1.4

Participation by Cluster by Tier by Domain, S502 Online

Cluster			Domain	
			Writing	Speaking
1	Tier	Pre-A	-	5,818
		A	115,739	59,302
		BC	21,534	72,159
	Total		137,273	137,273
2–3	Tier	Pre-A		9,382
		A	73,950	73,222
		BC	187,444	178,814
	Total		261,394	261,418
4–5	Tier	Pre-A		3,333
		A	47,588	31,571
		BC	194,323	207,011
	Total		241,911	241,915
6–8	Tier	Pre-A		5,915
		A	91,432	49,790
		BC	163,306	199,034
	Total		254,738	254,739
9–12	Tier	Pre-A		12,283
		A	65,351	76,897
		BC	143,879	120,045
	Total		209,230	209,225

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 1st grade and the smallest was 12th grade. Table 1.1.2.4 presents the percentages between Tier A and B/C and indicates that 4th grade had the smallest Tier A percentage and the highest Tier B/C percentage.

Table 1.1.2.1

Participation by Grade by State, S502 Online

State	Grade												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
AK	493	511	452	521	541	498	343	385	329	261	234	209	4,777
AL	3,373	3,534	3,789	3,570	3,038	2,676	2,497	1,966	1,463	1,285	1,016	787	28,994
BI	27	35	47	51	46	42	48	44	1	2	.	3	346
CO	7,847	8,025	7,697	6,900	5,658	4,399	4,338	4,178	3,536	2,827	2,355	1,937	59,697
DC	10	7	4	2	8	4	1	2	1	.	.	.	39
DD	743	793	759	713	562	478	340	331	208	151	123	67	5,268
DE	1,354	1,342	1,417	1,306	1,072	795	805	612	480	373	266	205	10,027
GA	12,301	12,558	12,452	13,371	10,244	7,597	6,859	5,685	5,310	4,201	2,920	2,227	95,725
HI	1,547	1,750	1,728	1,671	1,479	1,108	1,230	913	729	691	504	485	13,835
ID	1,893	2,125	1,972	1,951	1,368	1,039	1,442	1,203	971	895	713	494	16,066
IL	14,169	6,869	6,606	14,411	11,020	8,887	8,625	7,204	5,633	4,861	3,752	2,926	94,963
IN	7,695	7,558	7,949	7,602	6,291	5,378	5,165	3,990	3,114	3,023	3,302	2,542	63,609
KY	3,781	3,550	2,835	2,891	2,010	1,560	1,497	1,472	1,465	1,396	996	675	24,128
MA	9,029	8,784	7,554	6,532	4,638	3,756	4,068	3,972	3,620	3,491	2,543	1,890	59,877
MD	666	752	796	726	1,032	507	431	590	303	247	172	83	6,305
ME	248	273	248	212	179	180	154	168	133	133	112	106	2,146
MI	6,376	6,641	7,053	6,912	5,171	4,177	4,816	4,643	4,436	4,056	3,195	2,778	60,254
MN	5,866	5,954	5,844	5,397	3,710	2,579	2,736	2,601	2,232	1,994	1,603	1,148	41,664
MO	3,760	3,747	3,629	3,386	2,465	1,825	1,883	1,776	1,636	1,420	1,148	858	27,533
MP	47	94	102	152	135	4	.	1	30	4	.	.	569
MT	247	259	225	270	243	241	241	222	126	92	84	68	2,318
NC	10,482	11,159	11,739	11,376	9,694	7,888	8,281	6,221	4,988	4,233	3,317	2,454	91,832
ND	383	331	408	360	300	249	276	200	215	185	188	160	3,255
NH	407	427	444	411	296	206	248	291	217	227	158	126	3,458
NJ	7,635	7,212	7,066	6,156	4,611	3,329	3,100	3,138	2,819	2,990	2,554	1,826	52,436
NM	212	241	382	513	513	456	352	330	222	144	115	50	3,530
NV	2,254	2,225	2,241	4,730	3,274	2,222	2,372	2,530	1,932	1,963	1,501	1,310	28,554
OK	6,126	6,307	6,370	6,317	5,098	4,277	4,011	3,568	2,748	2,201	2,036	1,649	50,708
PA	3,117	3,297	3,320	3,367	2,844	2,919	2,853	2,686	2,423	2,176	1,889	1,570	32,461
RI	1,318	1,244	1,245	1,216	1,176	965	1,026	944	874	918	766	667	12,359
SC	3,467	3,656	3,837	3,970	3,494	2,808	3,247	3,143	3,478	2,804	2,185	1,685	37,774
SD	723	681	658	638	519	417	481	368	331	224	178	179	5,397
TN	6,031	5,760	5,536	4,981	3,601	2,831	2,828	2,749	2,660	2,555	1,959	1,295	42,786
UT	4,317	4,804	5,177	5,484	4,912	4,542	4,291	3,649	2,656	2,018	1,887	1,501	45,238
VA	4,917	4,036	3,720	2,424	1,740	4,158	4,066	3,750	2,710	2,852	2,419	1,347	38,139
VT	149	138	171	161	138	99	99	82	81	75	71	84	1,348
WI	4,044	4,182	4,557	4,272	3,440	2,883	2,868	2,834	2,028	1,604	1,372	1,078	35,162
WY	238	271	277	278	187	136	145	163	139	137	88	107	2,166
Total	137,292	131,132	130,306	135,201	106,747	88,115	88,063	78,604	66,277	58,709	47,721	36,576	1,104,743

Table 1.1.2.2

Participation by Grade by Gender, S502 Online

Grade		Gender			Total
		F	M	Missing	
1	Count	63,966	71,773	1,553	137,292
	% within Grade	46.59%	52.28%	1.13%	100.0%
2	Count	60,841	68,951	1,340	131,132
	% within Grade	46.40%	52.58%	1.02%	100.0%
3	Count	60,282	68,735	1,289	130,306
	% within Grade	46.26%	52.75%	0.99%	100.0%
4	Count	61,544	72,269	1,388	135,201
	% within Grade	45.52%	53.45%	1.03%	100.0%
5	Count	47,111	58,445	1,191	106,747
	% within Grade	44.13%	54.75%	1.12%	100.0%
6	Count	37,985	48,799	1,331	88,115
	% within Grade	43.11%	55.38%	1.51%	100.0%
7	Count	37,941	48,994	1,128	88,063
	% within Grade	43.08%	55.64%	1.28%	100.0
8	Count	33,382	44,269	953	78,604
	% within Grade	42.47%	56.32%	1.21%	100.0%
9	Count	28,032	37,136	1,109	66,277
	% within Grade	42.30%	56.03%	1.67%	100.0%
10	Count	25,337	32,614	758	58,709
	% within Grade	43.16%	55.55%	1.29%	100.0%
11	Count	21,081	25,990	650	47,721
	% within Grade	44.18%	54.46%	1.36%	100.0%
12	Count	16,462	19,735	379	36,576
	% within Grade	45.01%	53.96%	1.04%	100.0%
Total	Count	493,964	597,710	13,069	1,104,743
	% within Grade	44.71%	54.10%	1.18%	100.0%

Table 1.1.2.3

Participation by Grade by Ethnicity, S502 Online

Grade		Ethnicity			Total
		Hispanic	Non-Hispanic	Unknown	
1	Count	86,941	42,148	8,203	137,292
	% within Grade	63.33%	30.70%	5.97%	100.0%
2	Count	84,106	39,104	7,922	131,132
	% within Grade	64.14%	29.82%	6.04%	100.0%
3	Count	84,917	37,424	7,965	130,306
	% within Grade	65.17%	28.72%	6.11%	100.0%
4	Count	89,341	35,724	10,136	135,201
	% within Grade	66.08%	26.42%	7.50%	100.0%
5	Count	72,551	25,590	8,606	106,747
	% within Grade	67.97%	23.97%	8.06%	100.0%
6	Count	60,340	19,706	8,069	88,115
	% within Grade	68.48%	22.36%	9.16%	100.0%
7	Count	59,939	20,449	7,675	88,063
	% within Grade	68.06%	23.22%	8.72%	100.0%
8	Count	52,792	18,814	6,998	78,604
	% within Grade	67.16%	23.94%	8.90%	100.0%
9	Count	43,606	16,006	6,665	66,277
	% within Grade	65.79%	24.15%	10.06%	100.0%
10	Count	38,842	14,577	5,290	58,709
	% within Grade	66.16%	24.83%	9.01%	100.0%
11	Count	30,716	12,629	4,376	47,721
	% within Grade	64.37%	26.46%	9.17%	100.0%
12	Count	22,785	10,632	3,159	36,576
	% within Grade	62.29%	29.07%	8.64%	100.0%
Total	Count	726,876	292,803	85,064	1,104,743
	% within Grade	65.80%	26.50%	7.70%	100.0%

Table 1.1.2.4

Participation by Grade by Tier by Domain, S502 Online

Grade			Domain	
			Writing	Speaking
1	Tier	Pre-A	-	5,818
		A	115,739	59,302
		BC	21,534	72,159
	Total		137,273	137,279
2	Tier	Pre-A	-	3,125
		A	41,956	37,491
		BC	89,151	90,508
	Total		131,107	131,124
3	Tier	Pre-A	-	6,257
		A	31,994	35,731
		BC	98,293	88,306
	Total		130,287	130,294
4	Tier	Pre-A	-	1,281
		A	24,391	17,332
		BC	110,792	116,573
	Total		135,183	135,186
5	Tier	Pre-A	-	2,052
		A	23,197	14,239
		BC	83,531	90,438
	Total		106,728	106,729
6	Tier	Pre-A	-	1,311
		A	26,779	14,033
		BC	61,322	72,757
	Total		88,101	88,101
7	Tier	Pre-A	-	1,705
		A	32,084	12,580
		BC	55,968	73,767
	Total		88,052	88,052
8	Tier	Pre-A	-	2,899
		A	32,569	23,177
		BC	46,016	52,510
	Total		78,585	78,586

Grade			Domain	
			Writing	Speaking
9	Tier	Pre-A	-	2,442
		A	22,263	32,009
		BC	44,001	31,812
	Total		66,264	66,263
10	Tier	Pre-A	-	4,038
		A	19,579	21,399
		BC	39,114	33,256
	Total		58,693	58,693
11	Tier	Pre-A	-	3,198
		A	14,153	8,495
		BC	33,555	36,013
	Total		47,708	47,706
12	Tier	Pre-A	-	2,605
		A	9,356	14,994
		BC	27,209	18,964
	Total		36,565	36,563

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 in all clusters. Cluster 4–5 showed the highest average scale scores in all single domains across all clusters, and scores dropped in Cluster 6–8.

Table 1.2.1.2 demonstrates that female groups performed better than male groups in general except Clusters 4–12 in Listening. 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. Additional tables show this information by gender, and by race and ethnicity.

Table 1.2.1.1
Mean Scale Scores by Cluster, S502 Online

Cluster		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
1	Mean	315.14	287.65	243.68	248.36	281.97	265.79	295.97	270.55
	N	128,188	131,078	137,173	128,414	120,783	131,049	123,489	116,586
2-3	Mean	331.11	321.77	296.63	269.68	300.61	309.25	324.63	306.56
	N	244,652	246,934	261,229	243,666	229,501	246,854	233,167	219,334
4-5	Mean	408.9	349.6	333.77	303.3	356.4	341.76	367.5	346.12
	N	225,824	225,166	234,065	225,512	212,053	219,258	212,922	196,377
6-8	Mean	400.82	350.4	321.24	312.43	356.87	335.91	365.75	342.10
	N	236,488	236,259	244,573	235,766	221,119	228,966	223,174	204,407
9-12	Mean	399.06	383.73	347.77	313.95	356.61	365.91	388.42	362.81
	N	195,881	193,768	202,099	195,109	184,127	188,742	184,195	170,352

Table 1.2.1.2
Mean Scale Scores by Gender, S502 Online

Cluster	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
1	F	Mean	319.32	289.22	248.17	255.47	287.57	268.80	298.35	274.34
		N	59,975	61,063	63,920	60,107	56,772	61,056	57,762	54,785
	M	Mean	311.49	286.27	239.81	242	276.95	263.17	293.88	267.17
		N	66,774	68,520	71,700	66,839	62,642	68,498	64,327	60,467
	Missing	Mean	311.01	286.95	237.59	246.25	279.18	262.42	294.11	267.68
		N	1,439	1,495	1,553	1,468	1,369	1,495	1,400	1,334
2-3	F	Mean	333.28	323.48	302.86	276.72	305.16	313.23	326.48	310.63
		N	113,452	114,188	121,024	113,287	106,803	114,153	107,924	101,863
	M	Mean	329.45	320.33	291.39	263.68	296.83	305.93	323.14	303.15
		N	128,705	130,223	137,578	127,891	120,330	130,180	122,832	115,178
	Missing	Mean	318.1	318.2	283.33	257.8	287.9	300.73	318	296.69
		N	2,495	2,523	2,627	2,488	2,368	2,521	2,411	2,293
4-5	F	Mean	407.9	352.2	339.95	306.8	357.6	346.18	369	349.49
		N	101,816	100,957	105,120	101,635	95,911	98,347	95,909	88,757
	M	Mean	409.93	347.52	328.90	300.63	355.62	338.27	366.36	343.48
		N	121,608	121,787	126,436	121,486	113,898	118,542	114,730	105,511
	Missing	Mean	399.27	344.16	320.37	292.37	346.02	332.29	361.14	336.42
		N	2,400	2,422	2,509	2,391	2,244	2,369	2,283	2,109
6-8	F	Mean	400.03	353.04	326.60	315.07	357.75	339.94	367.33	345.11
		N	102,082	101,408	104,833	101,280	95,529	98,275	96,410	88,387
	M	Mean	401.54	348.43	317.26	310.53	356.31	332.91	364.6	339.86
		N	131,335	131,755	136,441	131,362	122,739	127,679	123,897	113,398
	Missing	Mean	396.64	347.3	315.70	306.59	351.75	331.30	362.31	337.30
		N	3,071	3,096	3,299	3,124	2,851	3,012	2,867	2,622
9-12	F	Mean	398.49	386.41	352.55	315.53	357.1	369.67	390.12	365.53
		N	85,407	84,193	87,634	84,738	80,262	81,936	80,315	74,197
	M	Mean	399.77	381.78	344.35	313.05	356.53	363.19	387.27	360.93
		N	107,754	106,944	111,683	107,681	101,314	104,259	101,368	93,850
	Missing	Mean	388.54	377.38	334.06	300.5	344.29	355.81	380.61	352.17
		N	2,720	2,631	2,782	2,690	2,551	2,547	2,512	2,305

Table 1.2.1.3

Mean Scale Scores by Ethnicity, S502 Online

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
1	Non-Hispanic Asian	Mean	328.66	308.39	263.59	260.94	295.05	286.19	314.66	288.76
		N	16,778	17,105	17,973	16,927	15,913	17,103	16,131	15,331
	Non-Hispanic Pacific Islander	Mean	290.67	279.73	230.39	233.48	262.22	254.99	283.22	257.18
		N	1,219	1,244	1,301	1,226	1,147	1,244	1,172	1,107
	Non-Hispanic Black	Mean	315.84	290.95	243.28	262.42	289.35	267.23	298.60	273.79
		N	6,658	6,866	7,254	6,691	6,192	6,866	6,362	5,935
	Hispanic (of any Race)	Mean	310.65	281.92	237.99	243.13	277.08	260.10	290.59	265.09
		N	81,264	83,288	86,877	81,351	76,576	83,266	78,489	74,086
	Non-Hispanic American Indian	Mean	315.30	285.83	236.33	247.28	281.78	261.19	295.01	267.31
		N	573	585	607	573	544	583	557	529
	Non-Hispanic Multiracial	Mean	331.06	299.51	254.74	257.21	294.45	277.25	308.81	281.99
		N	814	804	869	792	750	804	763	713
	Non-Hispanic White	Mean	329.02	296.12	256.39	260.00	294.86	276.36	306.12	281.83
		N	13,175	13,363	14,098	13,246	12,452	13,362	12,601	11,933
	Unknown	Mean	310.96	286.17	240.37	245.19	278.24	263.24	293.54	267.60
		N	7,707	7,823	8,194	7,608	7,209	7,821	7,414	6,952
2-3	Non-Hispanic Asian	Mean	346.58	335.16	312.72	278.49	312.79	324.11	338.79	320.73
		N	30,111	30,345	32,025	30,034	28,416	30,342	28,806	27,265
	Non-Hispanic Pacific Islander	Mean	309.33	315.28	294.41	251.09	280.15	304.86	313.41	297.17
		N	2,656	2,704	2,842	2,661	2,500	2,704	2,544	2,404
	Non-Hispanic Black	Mean	332.20	322.37	295.12	279.98	306.20	308.74	325.42	307.91
		N	13,500	13,766	14,746	13,536	12,489	13,761	12,751	11,832
	Hispanic (of any Race)	Mean	326.66	318.34	292.61	266.17	296.61	305.53	320.87	302.72
		N	158,261	159,867	168,900	157,681	148,501	159,804	150,967	142,037
	Non-Hispanic American Indian	Mean	327.42	319.48	286.89	259.45	292.88	303.21	322.11	299.68
		N	937	927	983	921	878	926	889	838
	Non-Hispanic Multiracial	Mean	352.25	333.28	308.15	278.55	315.70	320.93	339.01	319.42
		N	1,472	1,477	1,586	1,458	1,364	1,476	1,383	1,286
	Non-Hispanic White	Mean	345.71	328.96	307.58	280.66	313.63	318.35	334.11	316.94
		N	22,749	22,781	24,278	22,652	21,379	22,776	21,528	20,294
	Unknown	Mean	325.89	319.94	291.43	266.15	296.17	305.55	321.64	302.38
		N	14,966	15,067	15,869	14,723	13,974	15,065	14,299	13,378

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
4-5	Non-Hispanic Asian	Mean	420.38	363.12	347.84	311.53	366.31	355.62	380.49	358.81
		N	22,040	21,937	22,718	21,955	20,777	21,398	20,912	19,375
	Non-Hispanic Pacific Islander	Mean	397.99	345.11	333.22	291.97	345.13	339.39	361.22	341.19
		N	2,496	2,469	2,548	2,460	2,331	2,388	2,341	2,132
	Non-Hispanic Black	Mean	410.19	347.97	331.53	312.59	361.63	339.79	366.77	346.27
		N	11,790	11,764	12,328	11,787	10,935	11,411	10,990	9,973
	Hispanic (of any Race)	Mean	406.98	347.35	331.92	301.13	354.40	339.71	365.35	344.07
		N	151,150	150,757	156,674	151,190	142,133	146,788	142,528	131,606
	Non-Hispanic American Indian	Mean	405.11	342.15	323.11	289.20	347.43	332.84	361.36	336.97
		N	956	972	1,006	953	881	953	905	825
	Non-Hispanic Multiracial	Mean	423.30	361.03	343.24	313.41	368.36	352.15	379.25	356.29
		N	1,067	1,043	1,102	1,053	995	1,017	990	910
	Non-Hispanic White	Mean	417.92	357.23	340.85	314.64	366.55	349.17	375.52	354.33
		N	18,814	18,722	19,429	18,751	17,628	18,171	17,690	16,280
Unknown	Mean	401.33	345.31	326.23	295.58	348.66	335.64	362.16	339.42	
	N	17,511	17,502	18,260	17,363	16,373	17,132	16,566	15,276	
6-8	Non-Hispanic Asian	Mean	412.49	365.06	330.79	324.87	369.00	348.04	379.58	354.19
		N	18,665	18,556	19,073	18,353	17,333	17,922	17,652	16,033
	Non-Hispanic Pacific Islander	Mean	394.26	348.17	320.00	305.79	350.79	334.17	362.29	339.16
		N	2,358	2,378	2,533	2,419	2,182	2,290	2,205	1,988
	Non-Hispanic Black	Mean	403.95	350.99	317.80	320.75	362.46	334.45	367.08	342.62
		N	13,264	13,239	13,801	13,281	12,322	12,769	12,399	11,221
	Hispanic (of any Race)	Mean	398.87	348.41	320.67	310.01	354.70	334.65	363.77	340.60
		N	160,751	160,835	166,382	160,646	150,697	156,058	151,941	139,594
	Non-Hispanic American Indian	Mean	401.75	348.17	319.62	307.40	354.74	334.03	364.20	339.83
		N	1,333	1,349	1,405	1,351	1,245	1,307	1,256	1,145
	Non-Hispanic Multiracial	Mean	415.52	361.54	325.84	325.05	370.50	343.78	378.07	351.64
		N	920	930	972	930	852	894	855	779
	Non-Hispanic White	Mean	409.40	356.88	326.92	323.50	366.78	342.02	372.92	349.34
		N	18,101	17,951	18,656	17,929	16,873	17,400	16,974	15,540
Unknown	Mean	396.06	346.63	314.61	305.84	351.06	330.58	361.63	336.51	
	N	21,096	21,021	21,751	20,857	19,615	20,326	19,892	18,107	
9-12	Non-Hispanic Asian	Mean	413.00	397.87	360.60	331.92	372.67	379.49	402.54	377.18
		N	17,373	16,967	17,785	17,169	16,205	16,437	16,161	14,806
	Non-Hispanic Pacific Islander	Mean	390.23	378.43	348.99	303.28	346.74	363.78	382.14	358.23
		N	1,811	1,788	1,861	1,800	1,706	1,737	1,709	1,575
	Non-Hispanic Black	Mean	399.66	385.04	343.04	319.64	359.73	364.25	389.60	362.67
		N	14,160	13,910	14,727	14,247	13,293	13,506	13,094	12,121
	Hispanic (of any Race)	Mean	396.73	381.55	347.32	310.82	353.93	364.60	386.21	361.14
		N	127,451	126,535	131,651	127,106	120,039	123,435	120,314	111,530
	Non-Hispanic American Indian	Mean	404.16	384.72	345.80	314.95	359.70	365.45	390.67	363.00
		N	1,008	990	1,037	1,002	958	973	956	899
	Non-Hispanic Multiracial	Mean	414.31	394.42	355.33	330.33	372.59	375.42	400.68	374.90
		N	661	655	689	656	619	639	618	576
	Non-Hispanic White	Mean	411.36	390.60	351.94	325.52	368.39	371.35	396.81	370.00
		N	15,183	15,128	15,623	15,131	14,333	14,731	14,401	13,351
Unknown	Mean	391.39	379.00	338.65	305.06	348.08	358.84	382.74	355.04	
	N	18,234	17,795	18,726	17,998	16,974	17,284	16,942	15,494	

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. 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, S502 Online

Grade		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
1	Mean	315.14	287.65	243.68	248.36	281.97	265.79	295.97	270.55
	N	128,188	131,078	137,173	128,414	120,783	131,049	123,489	116,586
2	Mean	319.03	316.88	285.03	261.3	290.35	300.95	317.54	297.65
	N	122,368	123,835	131,018	121,582	114,298	123,791	116,639	109,241
3	Mean	343.21	326.68	308.30	278.03	310.8	317.61	331.73	315.40
	N	122,284	123,099	130,211	122,084	115,203	123,063	116,528	110,093
4	Mean	406.99	347.72	329.80	303.64	355.69	338.84	365.61	343.86
	N	126,013	125,652	130,725	125,763	118,071	122,315	118,624	109,168
5	Mean	411.31	351.97	338.79	302.98	357.39	345.44	369.9	348.95
	N	99,811	99,514	103,340	99,749	93,982	96,943	94,298	87,209
6	Mean	395.75	342.61	313.50	308.69	352.46	328.17	358.73	335.38
	N	81,591	81,786	84,864	81,256	76,048	79,491	76,997	70,490
7	Mean	401.23	351.99	322.69	313.19	357.49	337.46	367	343.43
	N	81,572	81,424	84,415	81,386	76,214	78,845	76,867	70,293
8	Mean	406.01	357.34	328.34	315.74	361.06	342.89	372.15	348.08
	N	73,325	73,049	75,294	73,124	68,857	70,630	69,310	63,624
9	Mean	395.17	379.06	343.59	310.32	352.89	361.39	383.96	358.53
	N	61,624	61,038	63,905	61,708	57,868	59,411	57,710	53,305
10	Mean	396.48	382.01	345.20	310.99	353.86	363.77	386.43	360.48
	N	55,044	54,455	56,707	54,818	51,854	53,046	51,830	48,010
11	Mean	403.3	387.6	351.91	318.2	360.8	369.98	392.4	366.91
	N	44,872	44,266	46,140	44,337	42,004	43,143	42,228	38,947
12	Mean	404.63	389.83	354.02	319.73	362.17	372.13	394.38	368.83
	N	34,341	34,009	35,347	34,246	32,401	33,142	32,427	30,090

Table 1.2.2.2

Mean Scale Scores by Grade by Gender, S502 Online

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
1	F	Mean	319.32	289.22	248.17	255.47	287.57	268.80	298.35	274.34
		N	59,975	61,063	63,920	60,107	56,772	61,056	57,762	54,785
	M	Mean	311.49	286.27	239.81	242.00	276.95	263.17	293.88	267.17
		N	66,774	68,520	71,700	66,839	62,642	68,498	64,327	60,467
	Missing	Mean	311.01	286.95	237.59	246.25	279.18	262.42	294.11	267.68
		N	1,439	1,495	1,553	1,468	1,369	1,495	1,400	1,334
2	F	Mean	321.85	318.18	291.24	268.48	295.28	304.70	319.27	301.67
		N	56,841	57,283	60,788	56,643	53,326	57,264	54,038	50,841
	M	Mean	316.70	315.78	279.80	255.09	286.12	297.80	316.08	294.24
		N	64,257	65,261	68,891	63,678	59,774	65,237	61,370	57,239
	Missing	Mean	310.50	315.09	272.03	252.57	281.80	293.36	313.67	289.83
		N	1,270	1,291	1,339	1,261	1,198	1,290	1,231	1,161
3	F	Mean	344.76	328.83	314.60	284.96	315.00	321.82	333.71	319.55
		N	56,611	56,905	60,236	56,644	53,477	56,889	53,886	51,022
	M	Mean	342.17	324.90	303.02	272.20	307.41	314.09	330.18	311.96
		N	64,448	64,962	68,687	64,213	60,556	64,943	61,462	57,939
	Missing	Mean	325.99	321.52	295.07	263.33	294.23	308.45	322.53	303.73
		N	1,225	1,232	1,288	1,227	1,170	1,231	1,180	1,132
4	F	Mean	406.19	350.20	335.93	307.65	357.24	343.15	367.07	347.25
		N	57,557	57,103	59,496	57,471	54,120	55,596	54,122	49,979
	M	Mean	407.90	345.73	324.85	300.39	354.56	335.37	364.51	341.15
		N	67,152	67,243	69,875	67,008	62,733	65,439	63,258	58,042
	Missing	Mean	395.95	341.86	316.11	293.16	344.66	328.84	358.32	333.29
		N	1,304	1,306	1,354	1,284	1,218	1,280	1,244	1,147
5	F	Mean	410.13	354.89	345.20	305.80	358.19	350.12	371.57	352.38
		N	44,259	43,854	45,624	44,164	41,791	42,751	41,787	38,778
	M	Mean	412.43	349.73	333.90	300.93	356.93	341.86	368.64	346.32
		N	54,456	54,544	56,561	54,478	51,165	53,103	51,472	47,469
	Missing	Mean	403.21	346.85	325.37	291.46	347.64	336.35	364.52	340.14
		N	1,096	1,116	1,155	1,107	1,026	1,089	1,039	962

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
6	F	Mean	394.83	344.90	319.02	310.77	353.00	332.07	360.04	338.20
		N	35,399	35,239	36,552	35,113	33,046	34,264	33,386	30,611
	M	Mean	396.48	340.83	309.24	307.14	352.08	325.15	357.70	333.18
		N	45,017	45,352	47,025	44,915	41,901	44,066	42,506	38,865
	Missing	Mean	395.48	342.91	312.13	305.51	350.95	327.34	358.70	334.68
N		1,175	1,195	1,287	1,228	1,101	1,161	1,105	1,014	
7	F	Mean	400.63	355.02	328.33	315.92	358.52	341.87	368.91	346.77
		N	35,323	35,134	36,353	35,116	33,040	34,021	33,369	30,560
	M	Mean	401.85	349.77	318.52	311.23	356.83	334.21	365.65	340.96
		N	45,223	45,267	46,965	45,251	42,234	43,824	42,554	38,871
	Missing	Mean	394.42	346.05	314.36	306.36	350.86	329.80	360.74	336.12
N		1,026	1,023	1,097	1,019	940	1,000	944	862	
8	F	Mean	405.20	360.05	333.31	318.96	362.23	346.75	373.74	351.02
		N	31,360	31,035	31,928	31,051	29,443	29,990	29,655	27,216
	M	Mean	406.73	355.35	324.74	313.46	360.32	340.08	371.01	345.95
		N	41,095	41,136	42,451	41,196	38,604	39,789	38,837	35,662
	Missing	Mean	400.82	354.73	322.33	308.38	353.85	338.48	369.00	342.23
N		870	878	915	877	810	851	818	746	
9	F	Mean	394.92	382.09	349.09	312.22	353.71	365.71	386.06	361.77
		N	26,186	25,804	26,948	26,136	24,628	25,072	24,507	22,618
	M	Mean	395.80	377.00	340.00	309.42	352.76	358.51	382.67	356.49
		N	34,418	34,223	35,896	34,544	32,282	33,364	32,253	29,821
	Missing	Mean	380.21	371.48	325.66	292.27	335.91	348.49	373.83	344.18
N		1,020	1,011	1,061	1,028	958	975	950	866	
10	F	Mean	395.52	384.63	349.88	312.50	354.04	367.45	387.93	363.01
		N	23,840	23,553	24,475	23,664	22,452	22,961	22,493	20,846
	M	Mean	397.35	380.02	341.84	310.05	353.91	361.05	385.34	358.67
		N	30,481	30,222	31,506	30,446	28,722	29,428	28,678	26,552
	Missing	Mean	391.20	379.39	333.77	300.52	345.54	356.58	382.54	352.67
N		723	680	726	708	680	657	659	612	
11	F	Mean	402.22	389.90	355.80	319.26	360.84	373.08	393.64	369.02
		N	19,879	19,543	20,342	19,543	18,567	19,024	18,683	17,185
	M	Mean	404.39	385.89	348.99	317.58	361.07	367.64	391.53	365.37
		N	24,377	24,131	25,168	24,195	22,864	23,541	22,976	21,240
	Missing	Mean	395.20	382.67	343.20	308.99	351.57	363.53	386.59	360.27
N		616	592	630	599	573	578	569	522	
12	F	Mean	404.30	391.99	358.40	321.05	362.73	375.43	395.80	371.27
		N	15,502	15,293	15,869	15,395	14,615	14,879	14,632	13,548
	M	Mean	405.09	388.19	350.58	318.81	361.88	369.56	393.36	366.95
		N	18,478	18,368	19,113	18,496	17,446	17,926	17,461	16,237
	Missing	Mean	395.39	381.56	343.26	309.98	353.15	362.26	385.89	360.00
N		361	348	365	355	340	337	334	305	

Table 1.2.2.3

Mean Scale Scores by Grade by Ethnicity, S502 Online

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
1	Non-Hispanic Asian	Mean	328.66	308.39	263.59	260.94	295.05	286.19	314.66	288.76
		N	16,778	17,105	17,973	16,927	15,913	17,103	16,131	15,331
	Non-Hispanic Pacific Islander	Mean	290.67	279.73	230.39	233.48	262.22	254.99	283.22	257.18
		N	1,219	1,244	1,301	1,226	1,147	1,244	1,172	1,107
	Non-Hispanic Black	Mean	315.84	290.95	243.28	262.42	289.35	267.23	298.6	273.79
		N	6,658	6,866	7,254	6,691	6,192	6,866	6,362	5,935
	Hispanic (of any Race)	Mean	310.65	281.92	237.99	243.13	277.08	260.10	290.59	265.09
		N	81,264	83,288	86,877	81,351	76,576	83,266	78,489	74,086
	Non-Hispanic American Indian	Mean	315.3	285.83	236.33	247.28	281.78	261.19	295.01	267.31
		N	573	585	607	573	544	583	557	529
	Non-Hispanic Multiracial	Mean	331.06	299.51	254.74	257.21	294.45	277.25	308.81	281.99
		N	814	804	869	792	750	804	763	713
	Non-Hispanic White	Mean	329.02	296.12	256.39	260	294.86	276.36	306.12	281.83
		N	13,175	13,363	14,098	13,246	12,452	13,362	12,601	11,933
Unknown	Mean	310.96	286.17	240.37	245.19	278.24	263.24	293.54	267.60	
	N	7,707	7,823	8,194	7,608	7,209	7,821	7,414	6,952	
2	Non-Hispanic Asian	Mean	335.79	328.93	303.61	270.65	303.43	316.36	331.14	312.45
		N	15,437	15,589	16,477	15,380	14,511	15,587	14,746	13,903
	Non-Hispanic Pacific Islander	Mean	297.51	311.94	281.37	242.95	270.2	296.68	307.57	288.90
		N	1,327	1,349	1,422	1,318	1,234	1,349	1,267	1,187
	Non-Hispanic Black	Mean	321.62	317.9	284.56	273.38	297.68	301.16	319.15	300.10
		N	6,815	6,962	7,434	6,775	6,259	6,960	6,451	5,940
	Hispanic (of any Race)	Mean	313.73	313.57	279.95	257.24	285.63	296.78	313.6	293.28
		N	78,489	79,602	84,041	78,072	73,366	79,565	74,954	70,240
	Non-Hispanic American Indian	Mean	314.79	313.61	274.07	250.77	282.08	293.61	314.02	289.71
		N	476	470	502	470	445	469	450	421
	Non-Hispanic Multiracial	Mean	343.31	326.17	298.60	268.73	306.18	312.21	331.65	310.67
		N	763	767	838	762	702	766	709	654
	Non-Hispanic White	Mean	335.15	323.45	298.04	273.05	304.57	310.69	326.99	308.84
		N	11,588	11,582	12,391	11,509	10,845	11,581	10,924	10,256
Unknown	Mean	314.31	316.02	280.28	258.51	286.48	298.00	315.43	294.11	
	N	7,473	7,514	7,913	7,296	6,936	7,514	7,138	6,640	

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall	
3	Non-Hispanic Asian	Mean	357.93	341.73	322.37	286.71	322.56	332.30	346.82	329.34	
		N	14,674	14,756	15,548	14,654	13,905	14,755	14,060	13,362	
	Non-Hispanic Pacific Islander	Mean	321.13	318.61	307.46	259.07	289.85	313.01	319.21	305.23	
		N	1,329	1,355	1,420	1,343	1,266	1,355	1,277	1,217	
	Non-Hispanic Black	Mean	342.99	326.94	305.86	286.59	314.75	316.50	331.83	315.78	
		N	6,685	6,804	7,312	6,761	6,230	6,801	6,300	5,892	
	Hispanic (of any Race)	Mean	339.39	323.06	305.14	274.93	307.33	314.21	328.03	311.95	
		N	79,772	80,265	84,859	79,609	75,135	80,239	76,013	71,797	
	Non-Hispanic American Indian	Mean	340.47	325.51	300.27	268.49	303.97	313.06	330.41	309.75	
		N	461	457	481	451	433	457	439	417	
	Non-Hispanic Multiracial	Mean	361.86	340.95	318.85	289.31	325.79	330.34	346.76	328.47	
		N	709	710	748	696	662	710	674	632	
	Non-Hispanic White	Mean	356.67	334.66	317.52	288.53	322.97	326.28	341.44	325.21	
		N	11,161	11,199	11,887	11,143	10,534	11,195	10,604	10,038	
	Unknown	Mean	337.43	323.84	302.51	273.65	305.71	313.07	327.84	310.53	
		N	7,493	7,553	7,956	7,427	7,038	7,551	7,161	6,738	
	4	Non-Hispanic Asian	Mean	419.99	361.8	344.96	312.19	366.51	353.59	379.48	357.50
			N	13,227	13,189	13,613	13,156	12,453	12,851	12,561	11,602
Non-Hispanic Pacific Islander		Mean	393.51	340.93	326.40	290.27	342.14	334.00	357.06	336.51	
		N	1,303	1,301	1,340	1,298	1,224	1,257	1,224	1,122	
Non-Hispanic Black		Mean	408.7	346.61	328.20	312.84	361.03	337.50	365.46	344.60	
		N	6,591	6,590	6,904	6,582	6,086	6,384	6,131	5,534	
Hispanic (of any Race)		Mean	404.57	345.1	327.39	301.13	353.22	336.30	363.01	341.31	
		N	83,295	83,001	86,420	83,244	78,138	80,801	78,349	72,197	
Non-Hispanic American Indian		Mean	402.18	339.63	316.38	289.37	346.43	328.32	359.23	334.18	
		N	528	543	555	531	491	533	507	467	
Non-Hispanic Multiracial		Mean	421.15	358.61	340.31	311.65	366.48	349.63	376.83	353.99	
		N	609	597	627	607	573	581	569	527	
Non-Hispanic White		Mean	416.42	355.61	337.78	315.18	366.18	346.84	374.01	352.63	
		N	11,010	10,971	11,393	10,952	10,274	10,644	10,347	9,475	
Unknown		Mean	399.23	343.44	322.50	296.12	348	332.85	360.15	337.32	
		N	9,450	9,460	9,873	9,393	8,832	9,264	8,936	8,244	

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
5	Non-Hispanic Asian	Mean	420.97	365.12	352.15	310.55	366	358.67	382.01	360.76
		N	8,813	8,748	9,105	8,799	8,324	8,547	8,351	7,773
	Non-Hispanic Pacific Islander	Mean	402.88	349.76	340.78	293.88	348.43	345.38	365.79	346.40
		N	1,193	1,168	1,208	1,162	1,107	1,131	1,117	1,010
	Non-Hispanic Black	Mean	412.08	349.71	335.76	312.28	362.37	342.70	368.42	348.34
		N	5,199	5,174	5,424	5,205	4,849	5,027	4,859	4,439
	Hispanic (of any Race)	Mean	409.94	350.11	337.49	301.14	355.84	343.89	368.21	347.42
		N	67,855	67,756	70,254	67,946	63,995	65,987	64,179	59,409
	Non-Hispanic American Indian	Mean	408.71	345.35	331.39	288.99	348.7	338.59	364.09	340.60
		N	428	429	451	422	390	420	398	358
	Non-Hispanic Multiracial	Mean	426.17	364.25	347.11	315.8	370.92	355.52	382.52	359.46
		N	458	446	475	446	422	436	421	383
	Non-Hispanic White	Mean	420.04	359.52	345.19	313.88	367.06	352.48	377.64	356.69
		N	7,804	7,751	8,036	7,799	7,354	7,527	7,343	6,805
	Unknown	Mean	403.79	347.52	330.64	294.95	349.43	338.92	364.52	341.90
		N	8,061	8,042	8,387	7,970	7,541	7,868	7,630	7,032
6	Non-Hispanic Asian	Mean	403.65	354.3	321.34	316.14	360.11	338.02	369.25	344.48
		N	6,225	6,219	6,448	6,175	5,790	6,044	5,866	5,361
	Non-Hispanic Pacific Islander	Mean	391.37	342.84	313.57	301.35	347.04	328.16	357.71	334.03
		N	868	864	919	883	805	833	813	731
	Non-Hispanic Black	Mean	398.43	342.69	309.03	316.15	357.6	325.95	359.66	335.30
		N	4,177	4,250	4,401	4,214	3,867	4,108	3,935	3,551
	Hispanic (of any Race)	Mean	394.35	340.93	313.06	306.9	350.89	327.13	357.14	334.22
		N	55,978	56,159	58,184	55,780	52,279	54,622	52,897	48,555
	Non-Hispanic American Indian	Mean	397.08	340.04	311.89	302.3	350.24	325.86	357.37	332.95
		N	452	452	479	453	421	442	422	388
	Non-Hispanic Multiracial	Mean	410.26	353.64	319.38	322.11	365.74	336.39	370.76	343.98
		N	355	358	374	347	322	346	331	298
	Non-Hispanic White	Mean	402.34	348.41	318.39	318.77	360.8	333.54	364.69	341.55
		N	6,080	6,017	6,315	6,025	5,655	5,858	5,684	5,206
	Unknown	Mean	392.53	340.39	308.55	304.12	348.39	324.44	356.22	331.53
		N	7,456	7,467	7,744	7,379	6,909	7,238	7,049	6,400

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
7	Non-Hispanic Asian	Mean	413.72	367.02	332.22	326.29	370.39	349.76	381.27	355.92
		N	6,513	6,453	6,626	6,382	6,046	6,218	6,156	5,584
	Non-Hispanic Pacific Islander	Mean	393.54	349.73	321.38	305.77	350.38	335.67	363.1	340.05
		N	838	857	915	861	775	826	784	710
	Non-Hispanic Black	Mean	403.28	352.19	318.76	321.06	361.99	335.51	367.66	343.20
		N	4,544	4,527	4,735	4,543	4,197	4,355	4,239	3,807
	Hispanic (of any Race)	Mean	399.25	350.09	322.25	310.76	355.29	336.31	365.07	341.97
		N	55,594	55,556	57,538	55,592	52,106	53,875	52,497	48,180
	Non-Hispanic American Indian	Mean	401.67	350.28	321.29	309.29	355.22	335.95	365.39	341.30
		N	459	467	487	461	420	453	431	384
	Non-Hispanic Multiracial	Mean	415.95	363.33	325.34	323.78	370.15	344.82	379.76	352.93
		N	298	298	324	305	276	291	271	249
	Non-Hispanic White	Mean	410.62	358.26	328.43	324.66	368.18	343.51	374.5	351.10
		N	6,235	6,217	6,453	6,205	5,801	6,014	5,845	5,336
	Unknown	Mean	396	347.46	315.17	306.01	351.22	331.21	362.1	337.02
		N	7,091	7,049	7,337	7,037	6,593	6,813	6,644	6,043
8	Non-Hispanic Asian	Mean	420.42	374.3	339.37	332.61	376.82	356.87	388.48	362.53
		N	5,927	5,884	5,999	5,796	5,497	5,660	5,630	5,088
	Non-Hispanic Pacific Islander	Mean	399.04	353.15	326.65	311.61	356.35	340.13	367.38	344.87
		N	652	657	699	675	602	631	608	547
	Non-Hispanic Black	Mean	409.68	357.67	325.09	324.73	367.33	341.47	373.42	348.77
		N	4,543	4,462	4,665	4,524	4,258	4,306	4,225	3,863
	Hispanic (of any Race)	Mean	403.59	355.07	327.60	312.67	358.33	341.39	369.82	346.27
		N	49,179	49,120	50,660	49,274	46,312	47,561	46,547	42,859
	Non-Hispanic American Indian	Mean	406.84	354.43	326.20	310.69	358.93	340.68	370.09	345.49
		N	422	430	439	437	404	412	403	373
	Non-Hispanic Multiracial	Mean	422.04	369.92	335.27	330.12	376.92	352.55	385.84	360.10
		N	267	274	274	278	254	257	253	232
	Non-Hispanic White	Mean	415.51	364.3	334.40	327.25	371.5	349.37	379.81	355.59
		N	5,786	5,717	5,888	5,699	5,417	5,528	5,445	4,998
	Unknown	Mean	400.13	352.91	321.01	307.63	353.89	336.97	367.27	341.61
		N	6,549	6,505	6,670	6,441	6,113	6,275	6,199	5,664

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
9	Non-Hispanic Asian	Mean	411.34	395.2	357.78	329.84	370.82	376.62	400.16	374.50
		N	4,885	4,730	4,990	4,843	4,569	4,574	4,513	4,135
	Non-Hispanic Pacific Islander	Mean	384.1	373.13	343.71	296.55	340.33	358.41	376.86	352.57
		N	647	636	663	640	607	617	606	552
	Non-Hispanic Black	Mean	398.62	381.86	340.86	318.2	358.4	361.28	387.04	360.15
		N	3,919	3,880	4,125	3,998	3,689	3,766	3,620	3,345
	Hispanic (of any Race)	Mean	393.13	377	343.33	307.8	350.65	360.26	381.9	357.13
		N	40,568	40,323	42,153	40,677	38,143	39,309	38,105	35,282
	Non-Hispanic American Indian	Mean	403.04	382.68	345.76	318.1	360.45	364.33	389.14	362.61
		N	320	316	331	321	303	308	303	281
	Non-Hispanic Multiracial	Mean	414.74	394.38	354.92	332.32	373.13	374.99	400.16	374.27
		N	247	252	257	243	230	247	239	222
	Non-Hispanic White	Mean	407.28	385.9	348.80	322.48	365.06	367.48	392.35	366.32
		N	4,860	4,835	4,996	4,835	4,577	4,707	4,599	4,260
	Unknown	Mean	384	372.72	331.40	297.08	340.41	351.84	376.06	347.66
		N	6,178	6,066	6,390	6,151	5,750	5,883	5,725	5,228
10	Non-Hispanic Asian	Mean	413.39	397.22	359.16	330.43	372.13	378.41	402.28	376.27
		N	4,664	4,558	4,767	4,583	4,346	4,427	4,363	4,005
	Non-Hispanic Pacific Islander	Mean	388.26	375.83	348.76	301.6	345.33	362.55	379.51	356.78
		N	473	472	496	483	451	460	447	420
	Non-Hispanic Black	Mean	398.82	384.28	341.12	318.55	358.9	363.04	388.87	361.62
		N	3,991	3,909	4,123	4,001	3,765	3,794	3,697	3,436
	Hispanic (of any Race)	Mean	393.06	379.26	344.09	307.08	350.23	361.83	383.48	358.07
		N	36,460	36,213	37,607	36,344	34,382	35,307	34,440	31,924
	Non-Hispanic American Indian	Mean	401.25	381.92	342.44	309.1	356.04	362.63	387.88	360.00
		N	283	277	291	284	273	272	269	256
	Non-Hispanic Multiracial	Mean	412.69	394.3	352.45	328.34	370.97	374.13	400.11	374.25
		N	176	169	177	175	167	160	160	146
	Non-Hispanic White	Mean	412.62	392.06	351.59	326.08	369.26	371.86	398.18	370.71
		N	4,015	4,016	4,150	4,035	3,813	3,910	3,815	3,555
	Unknown	Mean	390.68	378.26	338.01	303.64	346.97	358.20	381.94	354.12
		N	4,982	4,841	5,096	4,913	4,657	4,716	4,639	4,268

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre- hension	Overall
11	Non-Hispanic Asian	Mean	414.55	400.27	363.48	335.14	375.26	382.18	404.58	379.94
		N	4,227	4,144	4,342	4,173	3,940	4,025	3,928	3,608
	Non-Hispanic Pacific Islander	Mean	397.32	383.92	353.42	310.62	353.48	368.54	387.91	363.85
		N	383	378	390	375	357	365	363	328
	Non-Hispanic Black	Mean	401.79	387.17	345.61	321.97	361.98	366.68	391.62	365.05
		N	3,369	3,310	3,490	3,337	3,129	3,218	3,128	2,878
	Hispanic (of any Race)	Mean	401.18	385.55	351.68	315.01	358.19	368.84	390.34	365.33
		N	28,967	28,608	29,783	28,655	27,194	27,930	27,352	25,308
	Non-Hispanic American Indian	Mean	404.06	386.39	348.00	317.01	360.01	367.20	391.39	363.97
		N	241	239	250	234	223	237	229	212
	Non-Hispanic Multiracial	Mean	413.09	391.23	357.73	332.47	373.78	375.11	398.61	375.17
		N	134	132	146	131	121	130	120	110
	Non-Hispanic White	Mean	414.6	394.06	354.73	328.4	371.4	374.42	400.14	372.89
		N	3,449	3,450	3,546	3,417	3,239	3,359	3,286	3,032
	Unknown	Mean	398.58	384.41	344.31	311.87	355.23	364.58	388.9	361.46
		N	4,102	4,005	4,193	4,015	3,801	3,879	3,822	3,471
12	Non-Hispanic Asian	Mean	412.91	399.47	362.90	332.92	372.86	381.56	403.7	378.76
		N	3,597	3,535	3,686	3,570	3,350	3,411	3,357	3,058
	Non-Hispanic Pacific Islander	Mean	397.33	386.79	355.02	311.07	354.02	371.04	389.93	365.13
		N	308	302	312	302	291	295	293	275
	Non-Hispanic Black	Mean	399.76	388	345.68	320.44	360.08	367.17	391.71	364.79
		N	2,881	2,811	2,989	2,911	2,710	2,728	2,649	2,462
	Hispanic (of any Race)	Mean	403.74	388.66	354.56	317.31	360.63	371.78	393.31	368.15
		N	21,456	21,391	22,108	21,430	20,320	20,889	20,417	19,016
	Non-Hispanic American Indian	Mean	411.51	391.2	348.45	315.97	364.12	369.90	397.43	367.48
		N	164	158	165	163	159	156	155	150
	Non-Hispanic Multiracial	Mean	417.59	398.84	357.75	326.43	372.61	378.88	405.36	377.00
		N	104	102	109	107	101	102	99	98
	Non-Hispanic White	Mean	412.62	392.34	354.44	326.45	369.22	373.50	398.44	371.77
		N	2,859	2,827	2,931	2,844	2,704	2,755	2,701	2,504
	Unknown	Mean	398.03	385.95	347.16	314.92	356.05	366.67	389.43	363.06
		N	2,972	2,883	3,047	2,919	2,766	2,806	2,756	2,527

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 also correlated to Writing. In Clusters 4–5 and 6–8, Listening was correlated to Reading and Reading was correlated to 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, S502 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1.000	0.410	0.489	0.541
	N	128,188	123,489	128,158	120,783
Reading	Pearson Correlation		1.000	0.497	0.338
	N		131,078	131,049	123,167
Writing	Pearson Correlation			1.000	0.412
	N			137,173	128,385
Speaking	Pearson Correlation				1.000
	N				128,414

Table 1.2.3.2

Correlations Among Scale Scores: Grades 2–3, S502 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1.000	0.547	0.579	0.576
	N	244,652	233,167	244,575	229,501
Reading	Pearson Correlation		1.000	0.536	0.451
	N		246,934	246,854	231,428
Writing	Pearson Correlation			1.000	0.539
	N			261,229	243,589
Speaking	Pearson Correlation				1.000
	N				243,666

Table 1.2.3.3

Correlations Among Scale Scores: Grades 4–5, S502 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1.000	0.625	0.594	0.527
	N	225,824	212,922	219,425	212,053
Reading	Pearson Correlation		1.000	0.627	0.461
	N		225,166	219,258	211,584
Writing	Pearson Correlation			1.000	0.552
	N			234,065	219,193
Speaking	Pearson Correlation				1.000
	N				225,512

Table 1.2.3.4

Correlations Among Scale Scores: Grades 6–8, S502 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1.000	0.648	0.566	0.547
	N	236,488	223,174	228,622	221,119
Reading	Pearson Correlation		1.000	0.619	0.500
	N		236,259	228,966	220,983
Writing	Pearson Correlation			1.000	0.586
	N			244,573	227,975
Speaking	Pearson Correlation				1.000
	N				235,766

Table 1.2.3.5

Correlations Among Scale Scores: Grades 9–12, S502 Online

Domains	Pearson Correlations and N counts	Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1.000	0.702	0.496	0.597
	N	195,881	184,195	190,264	184,127
Reading	Pearson Correlation		1.000	0.532	0.600
	N		193,768	188,742	182,627
Writing	Pearson Correlation			1.000	0.582
	N			202,099	189,571
Speaking	Pearson Correlation				1.000
	N				195,109

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 (70%) in Proficiency Level (PL) 6, especially in Cluster 4–5. Clusters 1, 2–3, and 6–8 also had over 34% in PL 6. The Reading domain had 7% to 15% 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 2.7% in PL 5 and 6. In the Speaking domain, fewer than 1% were in PL 5 and PL 6; Cluster 4–5 showed 1.3% 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

Proficiency Level by Cluster (Count): Listening, S502 Online

Cluster	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	13,866	7,497	21,190	7,598	17,561	60,476	128,188
2–3	25,710	32,392	46,963	18,719	36,463	84,405	244,652
4–5	4,670	6,146	16,616	11,130	28,207	159,055	225,824
6–8	6,033	13,964	41,522	35,410	52,292	87,267	236,488
9–12	20,008	21,583	40,880	37,978	29,086	46,346	195,881

Table 1.3.1.1.2

Proficiency Level by Cluster (Percent): Listening, S502 Online

Cluster	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	10.8%	5.9%	16.5%	5.9%	13.7%	47.2%	100.0%
2–3	10.5%	13.2%	19.2%	7.7%	14.9%	34.5%	100.0%
4–5	2.1%	2.7%	7.4%	4.9%	12.5%	70.4%	100.0%
6–8	2.6%	5.9%	17.6%	15.0%	22.1%	36.9%	100.0%
9–12	10.2%	11.0%	20.9%	19.4%	14.9%	23.7%	100.0%

1.3.1.1.2 By Grade

Table 1.3.1.1.2.1

Proficiency Level by Grade (Count): Listening, S502 Online

Grade	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	13,866	7,497	21,190	7,598	17,561	60,476	128,188
2	13,328	17,275	26,363	9,632	14,881	40,889	122,368
3	12,382	15,117	20,600	9,087	21,582	43,516	122,284
4	2,146	3,097	7,892	5,349	12,920	94,609	126,013
5	2,524	3,049	8,724	5,781	15,287	64,446	99,811
6	1,407	3,878	12,890	10,591	22,508	30,317	81,591
7	1,739	4,946	15,596	12,812	16,008	30,471	81,572
8	2,887	5,140	13,036	12,007	13,776	26,479	73,325
9	3,823	7,978	13,049	11,807	8,447	16,520	61,624
10	6,396	5,472	12,229	10,273	7,907	12,767	55,044
11	5,253	4,194	8,632	9,126	8,132	9,535	44,872
12	4,536	3,939	6,970	6,772	4,600	7,524	34,341

Table 1.3.1.1.2.2

Proficiency Level by Grade (Percent): Listening, S502 Online

Grade	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	10.8%	5.9%	16.5%	5.9%	13.7%	47.2%	100.0%
2	10.9%	14.1%	21.5%	7.9%	12.2%	33.4%	100.0%
3	10.1%	12.4%	16.9%	7.4%	17.7%	35.6%	100.0%
4	1.7%	2.5%	6.3%	4.2%	10.3%	75.1%	100.0%
5	2.5%	3.1%	8.7%	5.8%	15.3%	64.6%	100.0%
6	1.7%	4.8%	15.8%	13.0%	27.6%	37.2%	100.0%
7	2.1%	6.1%	19.1%	15.7%	19.6%	37.4%	100.0%
8	3.9%	7.0%	17.8%	16.4%	18.8%	36.1%	100.0%
9	6.2%	13.0%	21.2%	19.2%	13.7%	26.8%	100.0%
10	11.6%	9.9%	22.2%	18.7%	14.4%	23.2%	100.0%
11	11.7%	9.4%	19.2%	20.3%	18.1%	21.3%	100.0%
12	13.2%	11.5%	20.3%	19.7%	13.4%	21.9%	100.0%

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, S502 Online

Cluster	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	33,633	37,404	22,139	12,675	11,764	13,463	131,078
2-3	31,734	74,928	54,831	24,793	38,111	22,537	246,934
4-5	33,644	49,714	40,450	28,361	42,027	30,970	225,166
6-8	71,912	62,126	53,493	10,688	21,125	16,915	236,259
9-12	32,082	53,239	37,088	12,889	29,255	29,215	193,768

Table 1.3.1.2.1.2

Proficiency Level by Cluster (Percent): Reading, S502 Online

Cluster	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	25.7%	28.5%	16.9%	9.7%	9.0%	10.3%	100.0%
2-3	12.9%	30.3%	22.2%	10.0%	15.4%	9.1%	100.0%
4-5	14.9%	22.1%	18.0%	12.6%	18.7%	13.8%	100.0%
6-8	30.4%	26.3%	22.6%	4.5%	8.9%	7.2%	100.0%
9-12	16.6%	27.5%	19.1%	6.7%	15.1%	15.1%	100.0%

1.3.1.2.2 By Grade

Table 1.3.1.2.2.1

Proficiency Level by Grade (Count): Reading, S502 Online

Grade	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	33,633	37,404	22,139	12,675	11,764	13,463	131,078
2	12,233	32,493	33,896	15,045	20,721	9,447	123,835
3	19,501	42,435	20,935	9,748	17,390	13,090	123,099
4	16,983	25,862	20,987	18,731	24,870	18,219	125,652
5	16,661	23,852	19,463	9,630	17,157	12,751	99,514
6	25,161	21,298	21,591	4,265	6,376	3,095	81,786
7	23,753	21,980	18,038	3,453	7,848	6,352	81,424
8	22,998	18,848	13,864	2,970	6,901	7,468	73,049
9	9,257	16,413	13,143	4,031	9,020	9,174	61,038
10	9,764	14,358	10,480	3,643	7,828	8,382	54,455
11	7,364	12,371	7,743	2,765	6,975	7,048	44,266
12	5,697	10,097	5,722	2,450	5,432	4,611	34,009

Table 1.3.1.2.2.2

Proficiency Level by Grade (Percent): Reading, S502 Online

Grade	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	25.7%	28.5%	16.9%	9.7%	9.0%	10.3%	100.0%
2	9.9%	26.2%	27.4%	12.2%	16.7%	7.6%	100.0%
3	15.8%	34.5%	17.0%	7.9%	14.1%	10.6%	100.0%
4	13.5%	20.6%	16.7%	14.9%	19.8%	14.5%	100.0%
5	16.7%	24.0%	19.6%	9.7%	17.2%	12.8%	100.0%
6	30.8%	26.0%	26.4%	5.2%	7.8%	3.8%	100.0%
7	29.2%	27.0%	22.2%	4.2%	9.6%	7.8%	100.0%
8	31.5%	25.8%	19.0%	4.1%	9.5%	10.2%	100.0%
9	15.2%	26.9%	21.5%	6.6%	14.8%	15.0%	100.0%
10	17.9%	26.4%	19.3%	6.7%	14.4%	15.4%	100.0%
11	16.6%	28.0%	17.5%	6.3%	15.8%	15.9%	100.0%
12	16.8%	29.7%	16.8%	7.2%	16.0%	13.6%	100.0%

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, S502 Online

Cluster	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	60,543	56,385	19,269	966	7	3	137,173
2–3	30,734	52,334	133,717	44,374	59	11	261,229
4–5	15,646	13,142	127,665	71,406	5,273	933	234,065
6–8	21,121	48,909	150,185	24,098	248	12	244,573
9–12	19,017	42,988	107,402	31,631	1,042	19	202,099

Table 1.3.1.3.1.2

Proficiency Level by Cluster (Percent): Writing, S502 Online

Cluster	Writing Proficiency Range						Total Total
	1	2	3	4	5	6	
1	44.1%	41.1%	14.1%	0.7%	0.0%	0.0%	100.0%
2–3	11.8%	20.0%	51.2%	17.0%	0.0%	0.0%	100.0%
4–5	6.7%	5.6%	54.5%	30.5%	2.3%	0.4%	100.0%
6–8	8.6%	20.0%	61.4%	9.9%	0.1%	0.0%	100.0%
9–12	9.4%	21.3%	53.1%	15.7%	0.5%	0.0%	100.0%

1.3.1.3.2 By Grade

Table 1.3.1.3.2.1

Proficiency Level by Grade (Count): Writing, S502 Online

Grade	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	60,543	56,385	19,269	966	7	3	137,173
2	20,521	34,735	63,983	11,773	5	1	131,018
3	10,213	17,599	69,734	32,601	54	10	130,211
4	9,366	5,449	79,081	34,522	1,695	612	130,725
5	6,280	7,693	48,584	36,884	3,578	321	103,340
6	6,091	15,867	56,899	5,935	69	3	84,864
7	6,751	20,334	45,359	11,899	70	2	84,415
8	8,279	12,708	47,927	6,264	109	7	75,294
9	4,490	13,289	32,261	13,608	245	12	63,905
10	4,850	10,832	34,470	6,191	361	3	56,707
11	5,897	10,753	22,717	6,406	366	1	46,140
12	3,780	8,114	17,954	5,426	70	3	35,347

Table 1.3.1.3.2.2

Proficiency Level by Grade (Percent): Writing, S502 Online

Grade	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	44.1%	41.1%	14.1%	0.7%	0.0%	0.0%	100.0%
2	15.7%	26.5%	48.8%	9.0%	0.0%	0.0%	100.0%
3	7.8%	13.5%	53.6%	25.0%	0.0%	0.0%	100.0%
4	7.2%	4.2%	60.5%	26.4%	1.3%	0.5%	100.0%
5	6.1%	7.4%	47.0%	35.7%	3.5%	0.3%	100.0%
6	7.2%	18.7%	67.1%	7.0%	0.1%	0.0%	100.0%
7	8.0%	24.1%	53.7%	14.1%	0.1%	0.0%	100.0%
8	11.0%	16.9%	63.7%	8.3%	0.1%	0.0%	100.0%
9	7.0%	20.8%	50.5%	21.3%	0.4%	0.0%	100.0%
10	8.6%	19.1%	60.8%	10.9%	0.6%	0.0%	100.0%
11	12.8%	23.3%	49.2%	13.9%	0.8%	0.0%	100.0%
12	10.7%	23.0%	50.8%	15.4%	0.2%	0.0%	100.0%

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, S502 Online

Cluster	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	21,906	52,175	37,696	15,807	752	78	128,414
2–3	46,436	76,033	93,241	25,868	1,644	444	243,666
4–5	28,853	62,975	89,718	40,858	2,897	211	225,512
6–8	46,436	68,740	92,243	27,724	556	67	235,766
9–12	61,022	56,444	68,626	8,743	221	53	195,109

Table 1.3.1.4.1.2

Proficiency Level by Cluster (Percent): Speaking, S502 Online

Cluster	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	17.1%	40.6%	29.4%	12.3%	0.6%	0.1%	100.0%
2–3	19.1%	31.2%	38.3%	10.6%	0.7%	0.2%	100.0%
4–5	12.8%	27.9%	39.8%	18.1%	1.3%	0.1%	100.0%
6–8	19.7%	29.2%	39.1%	11.8%	0.2%	0.0%	100.0%
9–12	31.3%	28.9%	35.2%	4.5%	0.1%	0.0%	100.0%

1.3.1.4.2 By Grade

Table 1.3.1.4.2.1

Proficiency Level by Grade (Count): Speaking, S502 Online

Grade	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	21,906	52,175	37,696	15,807	752	78	128,414
2	22,491	45,570	41,370	11,130	905	116	121,582
3	23,945	30,463	51,871	14,738	739	328	122,084
4	14,341	31,749	51,170	26,221	2,159	123	125,763
5	14,512	31,226	38,548	14,637	738	88	99,749
6	13,318	24,508	31,703	11,559	158	10	81,256
7	15,003	27,248	29,227	9,608	275	25	81,386
8	18,115	16,984	31,313	6,557	123	32	73,124
9	16,615	22,244	18,486	4,241	113	9	61,708
10	20,018	13,301	19,102	2,359	27	11	54,818
11	13,350	11,272	18,228	1,423	47	17	44,337
12	11,039	9,627	12,810	720	34	16	34,246

Table 1.3.1.4.2.2

Proficiency Level by Grade (Percent): Speaking, S502 Online

Grade	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	17.1%	40.6%	29.4%	12.3%	0.6%	0.1%	100.0%
2	18.5%	37.5%	34.0%	9.2%	0.7%	0.1%	100.0%
3	19.6%	25.0%	42.5%	12.1%	0.6%	0.3%	100.0%
4	11.4%	25.3%	40.7%	20.9%	1.7%	0.1%	100.0%
5	14.6%	31.3%	38.6%	14.7%	0.7%	0.1%	100.0%
6	16.4%	30.2%	39.0%	14.2%	0.2%	0.0%	100.0%
7	18.4%	33.5%	35.9%	11.8%	0.3%	0.0%	100.0%
8	24.8%	23.2%	42.8%	9.0%	0.2%	0.0%	100.0%
9	26.9%	36.1%	30.0%	6.9%	0.2%	0.0%	100.0%
10	36.5%	24.3%	34.9%	4.3%	0.1%	0.0%	100.0%
11	30.1%	25.4%	41.1%	3.2%	0.1%	0.0%	100.0%
12	32.2%	28.1%	37.4%	2.1%	0.1%	0.1%	100.0%

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, S502 Online

Cluster	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	14,073	20,921	39,274	27,990	16,225	2,300	120,783
2–3	27,277	49,643	69,944	62,246	18,680	1,711	229,501
4–5	8,496	15,532	49,128	84,741	44,066	10,090	212,053
6–8	15,709	31,011	79,693	75,842	16,485	2,379	221,119
9–12	32,710	35,664	71,072	38,688	5,161	832	184,127

Table 1.3.2.1.1.2

Proficiency Level by Cluster (Percent): Oral, S502 Online

Cluster	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	11.7%	17.3%	32.5%	23.2%	13.4%	1.9%	100.0%
2–3	11.9%	21.6%	30.5%	27.1%	8.1%	0.8%	100.0%
4–5	4.0%	7.3%	23.2%	40.0%	20.8%	4.8%	100.0%
6–8	7.1%	14.0%	36.0%	34.3%	7.5%	1.1%	100.0%
9–12	17.8%	19.4%	38.6%	21.0%	2.8%	0.5%	100.0%

1.3.2.1.2 By Grade

Table 1.3.2.1.2.1

Proficiency Level by Grade (Count): Oral, S502 Online

Grade	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	14,073	20,921	39,274	27,990	16,225	2,300	120,783
2	14,105	28,453	35,708	25,918	9,254	860	114,298
3	13,172	21,190	34,236	36,328	9,426	851	115,203
4	3,830	7,527	26,801	45,661	27,378	6,874	118,071
5	4,666	8,005	22,327	39,080	16,688	3,216	93,982
6	3,911	9,528	28,503	27,837	5,666	603	76,048
7	5,364	11,272	26,712	26,406	5,560	900	76,214
8	6,434	10,211	24,478	21,599	5,259	876	68,857
9	8,763	11,364	21,988	13,694	1,784	275	57,868
10	10,251	10,064	18,964	10,829	1,508	238	51,854
11	7,490	8,082	16,652	8,410	1,169	201	42,004
12	6,206	6,154	13,468	5,755	700	118	32,401

Table 1.3.2.1.2.2

Proficiency Level by Grade (Percent): Oral, S502 Online

Grade	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	11.7%	17.3%	32.5%	23.2%	13.4%	1.9%	100.0%
2	12.3%	24.9%	31.2%	22.7%	8.1%	0.8%	100.0%
3	11.4%	18.4%	29.7%	31.5%	8.2%	0.7%	100.0%
4	3.2%	6.4%	22.7%	38.7%	23.2%	5.8%	100.0%
5	5.0%	8.5%	23.8%	41.6%	17.8%	3.4%	100.0%
6	5.1%	12.5%	37.5%	36.6%	7.5%	0.8%	100.0%
7	7.0%	14.8%	35.1%	34.7%	7.3%	1.2%	100.0%
8	9.3%	14.8%	35.6%	31.4%	7.6%	1.3%	100.0%
9	15.1%	19.6%	38.0%	23.7%	3.1%	0.5%	100.0%
10	19.8%	19.4%	36.6%	20.9%	2.9%	0.5%	100.0%
11	17.8%	19.2%	39.6%	20.0%	2.8%	0.5%	100.0%
12	19.2%	19.0%	41.6%	17.8%	2.2%	0.4%	100.0%

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, S502 Online

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	39,550	55,256	28,131	6,422	1,466	224	131,049
2-3	25,109	56,157	118,803	43,045	3,539	201	246,854
4-5	19,761	22,349	93,013	66,696	14,248	3,191	219,258
6-8	34,352	60,497	105,799	25,649	2,488	181	228,966
9-12	18,969	45,123	84,155	33,987	6,206	302	188,742

Table 1.3.2.2.1.2

Proficiency Level by Cluster (Percent): Literacy, S502 Online

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	30.2%	42.2%	21.5%	4.9%	1.1%	0.2%	100.0%
2-3	10.2%	22.8%	48.1%	17.4%	1.4%	0.1%	100.0%
4-5	9.0%	10.2%	42.4%	30.4%	6.5%	1.5%	100.0%
6-8	15.0%	26.4%	46.2%	11.2%	1.1%	0.1%	100.0%
9-12	10.1%	23.9%	44.6%	18.0%	3.3%	0.2%	100.0%

1.3.2.2.2 By Grade

Table 1.3.2.2.2.1

Proficiency Level by Grade (Count): Literacy, S502 Online

Grade	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	39,550	55,256	28,131	6,422	1,466	224	131,049
2	13,435	32,321	59,107	17,734	1,117	77	123,791
3	11,674	23,836	59,696	25,311	2,422	124	123,063
4	10,902	11,847	52,768	37,605	7,536	1,657	122,315
5	8,859	10,502	40,245	29,091	6,712	1,534	96,943
6	11,206	22,340	39,313	6,181	398	53	79,491
7	11,279	20,331	37,080	9,100	976	79	78,845
8	11,867	17,826	29,406	10,368	1,114	49	70,630
9	5,294	11,836	27,890	12,062	2,193	136	59,411
10	5,717	12,472	23,476	9,423	1,849	109	53,046
11	4,183	11,215	18,741	7,548	1,415	41	43,143
12	3,775	9,600	14,048	4,954	749	16	33,142

Table 1.3.2.2.2.2

Proficiency Level by Grade (Percent): Literacy, S502 Online

Grade	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	30.2%	42.2%	21.5%	4.9%	1.1%	0.2%	100.0%
2	10.9%	26.1%	47.8%	14.3%	0.9%	0.1%	100.0%
3	9.5%	19.4%	48.5%	20.6%	2.0%	0.1%	100.0%
4	8.9%	9.7%	43.1%	30.7%	6.2%	1.4%	100.0%
5	9.1%	10.8%	41.5%	30.0%	6.9%	1.6%	100.0%
6	14.1%	28.1%	49.5%	7.8%	0.5%	0.1%	100.0%
7	14.3%	25.8%	47.0%	11.5%	1.2%	0.1%	100.0%
8	16.8%	25.2%	41.6%	14.7%	1.6%	0.1%	100.0%
9	8.9%	19.9%	46.9%	20.3%	3.7%	0.2%	100.0%
10	10.8%	23.5%	44.3%	17.8%	3.5%	0.2%	100.0%
11	9.7%	26.0%	43.4%	17.5%	3.3%	0.1%	100.0%
12	11.4%	29.0%	42.4%	15.0%	2.3%	0.1%	100.0%

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, S502 Online

Cluster	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	12,164	26,143	33,982	14,534	18,387	18,279	123,489
2–3	20,815	52,227	58,341	30,581	38,438	32,765	233,167
4–5	10,103	25,459	34,324	28,853	50,634	63,549	212,922
6–8	26,433	50,667	55,887	34,598	30,348	25,241	223,174
9–12	20,969	41,829	39,640	23,720	30,157	27,880	184,195

Table 1.3.2.3.1.2

Proficiency Level by Cluster (Percent): Comprehension, S502 Online

Cluster	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	9.9%	21.2%	27.5%	11.8%	14.9%	14.8%	100.0%
2–3	8.9%	22.4%	25.0%	13.1%	16.5%	14.1%	100.0%
4–5	4.7%	12.0%	16.1%	13.6%	23.8%	29.9%	100.0%
6–8	11.8%	22.7%	25.0%	15.5%	13.6%	11.3%	100.0%
9–12	11.4%	22.7%	21.5%	12.9%	16.4%	15.1%	100.0%

1.3.2.3.2 By Grade

Table 1.3.2.3.2.1

Proficiency Level by Grade (Count): Comprehension, S502 Online

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	12,164	26,143	33,982	14,534	18,387	18,279	123,489
2	7,725	26,571	31,020	17,046	19,381	14,896	116,639
3	13,090	25,656	27,321	13,535	19,057	17,869	116,528
4	4,098	13,675	18,695	15,301	29,166	37,689	118,624
5	6,005	11,784	15,629	13,552	21,468	25,860	94,298
6	7,257	18,576	21,109	13,143	10,962	5,950	76,997
7	9,234	16,678	19,273	12,152	10,103	9,427	76,867
8	9,942	15,413	15,505	9,303	9,283	9,864	69,310
9	5,345	12,674	12,996	7,893	9,731	9,071	57,710
10	6,349	11,926	11,116	6,472	8,059	7,908	51,830
11	5,065	9,605	8,756	5,248	6,893	6,661	42,228
12	4,210	7,624	6,772	4,107	5,474	4,240	32,427

Table 1.3.2.3.2.2

Proficiency Level by Grade (Percent): Comprehension, S502 Online

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	9.9%	21.2%	27.5%	11.8%	14.9%	14.8%	100.0%
2	6.6%	22.8%	26.6%	14.6%	16.6%	12.8%	100.0%
3	11.2%	22.0%	23.5%	11.6%	16.4%	15.3%	100.0%
4	3.5%	11.5%	15.8%	12.9%	24.6%	31.8%	100.0%
5	6.4%	12.5%	16.6%	14.4%	22.8%	27.4%	100.0%
6	9.4%	24.1%	27.4%	17.1%	14.2%	7.7%	100.0%
7	12.0%	21.7%	25.1%	15.8%	13.1%	12.3%	100.0%
8	14.3%	22.2%	22.4%	13.4%	13.4%	14.2%	100.0%
9	9.3%	22.0%	22.5%	13.7%	16.9%	15.7%	100.0%
10	12.3%	23.0%	21.5%	12.5%	15.6%	15.3%	100.0%
11	12.0%	22.8%	20.7%	12.4%	16.3%	15.8%	100.0%
12	13.0%	23.5%	20.9%	12.7%	16.9%	13.1%	100.0%

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, S502 Online

Cluster	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	18,792	45,534	41,277	8,829	1,939	215	116,586
2-3	21,200	49,971	98,627	45,261	4,167	108	219,334
4-5	11,901	18,946	70,305	75,158	17,615	2,452	196,377
6-8	20,794	43,729	101,979	34,730	2,969	206	204,407
9-12	20,959	37,055	77,093	31,175	3,899	171	170,352

Table 1.3.2.4.1.2

Proficiency Level by Cluster (Percent): Overall, S502 Online

Cluster	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	16.1%	39.1%	35.4%	7.6%	1.7%	0.2%	100.0%
2-3	9.7%	22.8%	45.0%	20.6%	1.9%	0.1%	100.0%
4-5	6.1%	9.7%	35.8%	38.3%	9.0%	1.3%	100.0%
6-8	10.2%	21.4%	49.9%	17.0%	1.5%	0.1%	100.0%
9-12	12.3%	21.8%	45.3%	18.3%	2.3%	0.1%	100.0%

1.3.2.4.2 By Grade

Table 1.3.2.4.2.1

Proficiency Level by Grade (Count): Overall, S502 Online

Grade	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	18,792	45,534	41,277	8,829	1,939	215	116,586
2	10,997	29,238	48,762	18,464	1,713	67	109,241
3	10,203	20,733	49,865	26,797	2,454	41	110,093
4	6,142	9,929	38,563	42,625	10,394	1,515	109,168
5	5,759	9,017	31,742	32,533	7,221	937	87,209
6	5,839	15,615	38,657	9,772	552	55	70,490
7	7,081	14,676	34,689	12,659	1,098	90	70,293
8	7,874	13,438	28,633	12,299	1,319	61	63,624
9	5,517	10,167	25,291	10,880	1,366	84	53,305
10	6,671	10,173	21,072	8,859	1,178	57	48,010
11	4,789	8,846	17,498	6,911	880	23	38,947
12	3,982	7,869	13,232	4,525	475	7	30,090

Table 1.3.2.4.2.2

Proficiency Level by Grade (Percent): Overall, S502 Online

Grade	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	16.1%	39.1%	35.4%	7.6%	1.7%	0.2%	100.0%
2	10.1%	26.8%	44.6%	16.9%	1.6%	0.1%	100.0%
3	9.3%	18.8%	45.3%	24.3%	2.2%	0.0%	100.0%
4	5.6%	9.1%	35.3%	39.1%	9.5%	1.4%	100.0%
5	6.6%	10.3%	36.4%	37.3%	8.3%	1.1%	100.0%
6	8.3%	22.2%	54.8%	13.9%	0.8%	0.1%	100.0%
7	10.1%	20.9%	49.4%	18.0%	1.6%	0.1%	100.0%
8	12.4%	21.1%	45.0%	19.3%	2.1%	0.1%	100.0%
9	10.4%	19.1%	47.5%	20.4%	2.6%	0.2%	100.0%
10	13.9%	21.2%	43.9%	18.5%	2.5%	0.1%	100.0%
11	12.3%	22.7%	44.9%	17.7%	2.3%	0.1%	100.0%
12	13.2%	26.2%	44.0%	15.0%	1.6%	0.0%	100.0%

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, the Center for Applied Linguistics (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 (CAL, 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 *Maintaining the ACCESS for ELLs Online Writing Scale: Preparations for the Series 501 redesign: Technical brief* CAL (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 (CAL, 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:

- Listening: (Ability Measure in Logits * 37.571) + 316.637
- Reading: (Ability Measure in Logits * 26.000) + 323.272
- Writing: (Ability Measure in Logits * 26.851) + 303.332

- Speaking: (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 (CAL, 2016). In the domains of Writing and Speaking, we conducted a study in the summer of 2016 to reconstruct the logit scale (CAL, 2017).

PL scores are interpretations of these scale scores in terms of the proficiency levels described in the WIDA English Language Development (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* (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 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 degrade¹ 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 followed conservative guidelines in the development of ACCESS for ELLs, the vast majority of dichotomous items on the test forms have mean square fit statistics in the range of 0.5 to 1.5; thus, they fit the range that is “productive for measurement” according to the guidelines above.

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 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). For dichotomously scored items (Listening and Reading), the fourth column shows the *p* value (percentage of correct answers on that item). The

¹ We interpret “degrade” here in the sense of lowering the quality of the measurement system.

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 all items and tasks have infit mean square statistics less than 2.0 for all grade clusters and domains, indicating that the items and tasks provide good measurement for students around the ability range that the items and tasks are targeting. As discussed earlier, the outfit mean square statistic is sensitive to outlier responses and ratings that are not close to the ability range that the items and tasks are targeting. There are three items in Listening grade-level Cluster 2–3 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.

2.2 DIF Analysis and Summary

Differential item functioning (DIF) analysis investigates 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. We exclude students for whom gender or ethnicity² 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 subsequent to 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 one group is favored; a negative result indicates that the other 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 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

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.

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 this, the absolute value is taken in JMetrik (Meyer, 2018), thereby restricting the range of the rescaled effect size (standardized P-DIF*) 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 versus female or Hispanic versus non-Hispanic other ethnicities. 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.

For all items, bias and sensitivity review occurs prior to any field testing (see Part 1 Section 2.3.3). If a task or item shows C-level (or CC-level) DIF, an additional bias review panel is convened.

Panel members are drawn from CAL staff members who have expertise in instruction and/or professional development for ELs. The panel includes a mix of women and men, as well as staff who have a language other than English as a first language, with attention to obtaining representation from Spanish and non-Spanish language backgrounds. The panel is asked to discuss the item and come to a consensus on whether they believe or do not believe that the item demonstrates 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 502, 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 the Listening Grade 2–3 item being reviewed in a previously held panel. 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.

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 as a whole 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 a number of 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 6 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 P: 1 tasks, as it is assumed that they would be able to get the full 2 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 2 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 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	115,655	0	12	5.03	2.57
Total	115,655	0	12	5.03	2.57

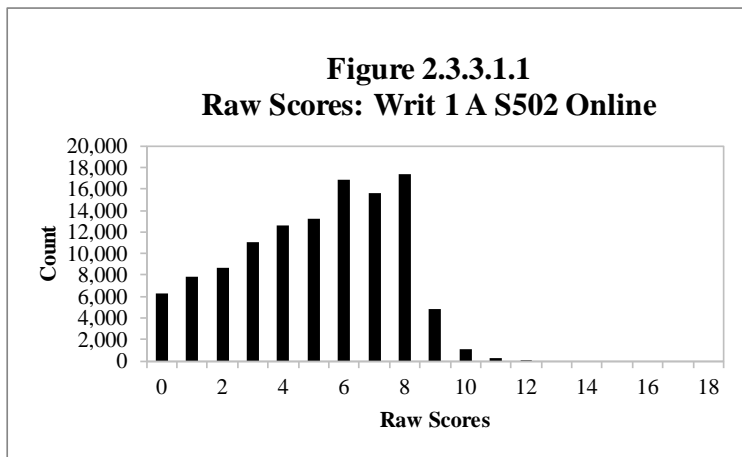
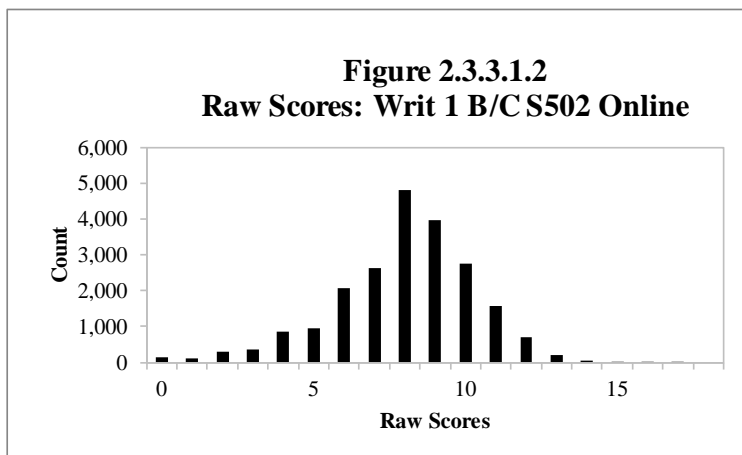


Table 2.3.3.1.2

Raw Score Descriptive Statistics: Writ 1 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	21,518	0	17	7.99	2.32
Total	21,518	0	17	7.99	2.32



2.3.3.2 Grades 2–3

Table 2.3.3.2.1

Raw Score Descriptive Statistics: Writ 2-3 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	41,920	0	15	6.05	3.24
3	31,972	0	15	7.04	3.23
Total	73,892	0	15	6.48	3.27

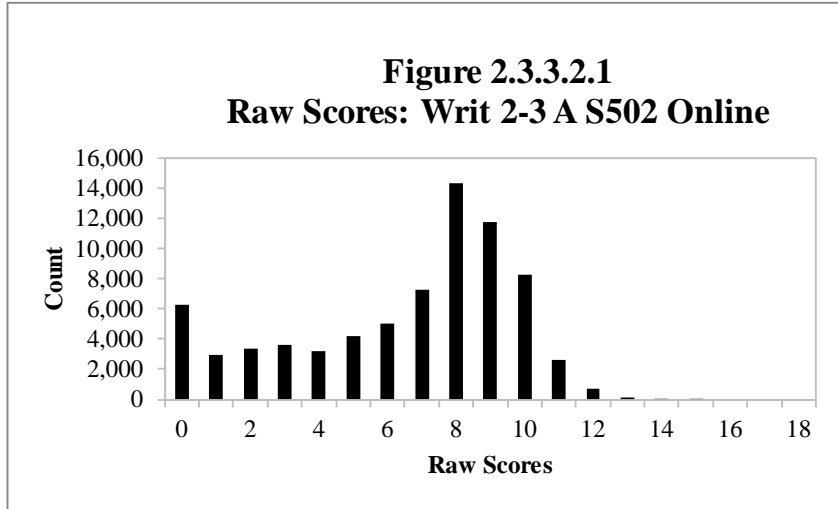
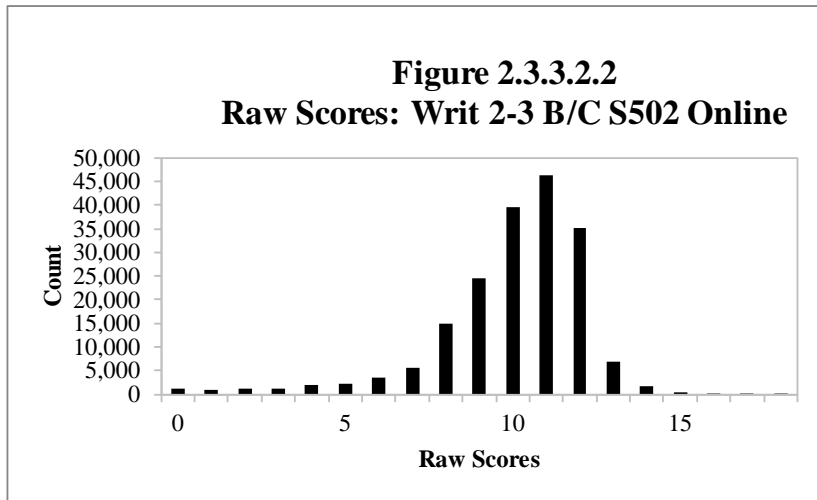


Table 2.3.3.2.2

Raw Score Descriptive Statistics: Writ 2-3 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	89,098	0	17	9.29	2.45
3	98,239	0	18	10.64	1.79
Total	187,337	0	18	10.00	2.23



2.3.3.3 Grades 4–5

Table 2.3.3.3.1

Raw Score Descriptive Statistics: Writ 4-5 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	23,547	0	15	5.38	3.14
5	22,475	0	15	6.18	3.13
Total	46,022	0	15	5.77	3.16

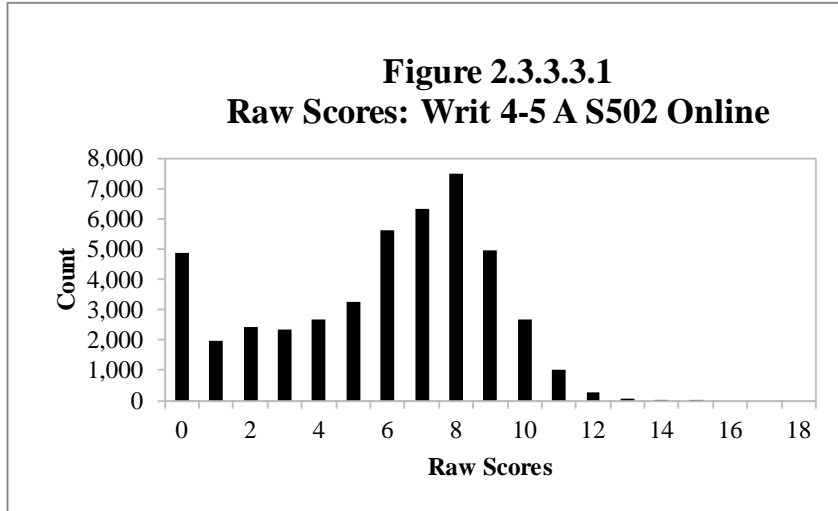
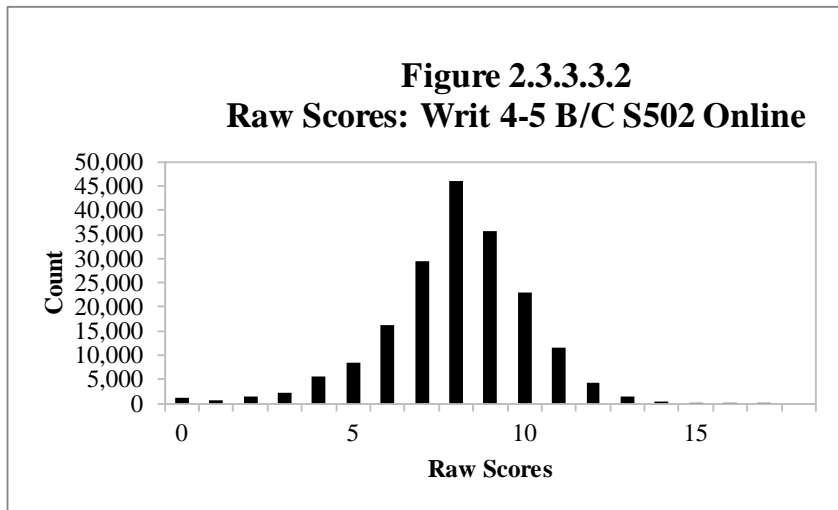


Table 2.3.3.3.2

Raw Score Descriptive Statistics: Writ 4-5 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	107,178	0	17	7.64	2.16
5	80,865	0	17	8.44	1.98
Total	188,043	0	17	7.98	2.12



2.3.3.4 Grades 6–8

Table 2.3.3.4.1

Raw Score Descriptive Statistics: Writ 6-8 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	26,020	0	15	6.64	2.79
7	31,064	0	15	7.30	2.77
8	31,503	16	7.73	2.76	
Total	88,587	0	16	7.26	2.80

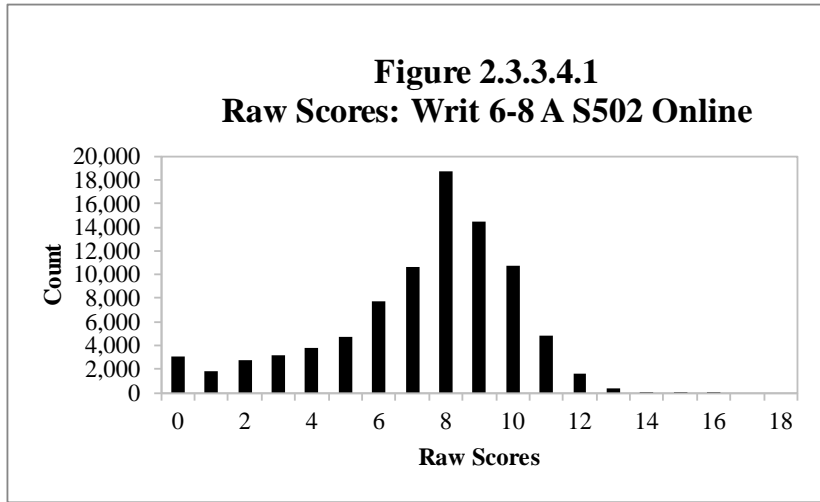
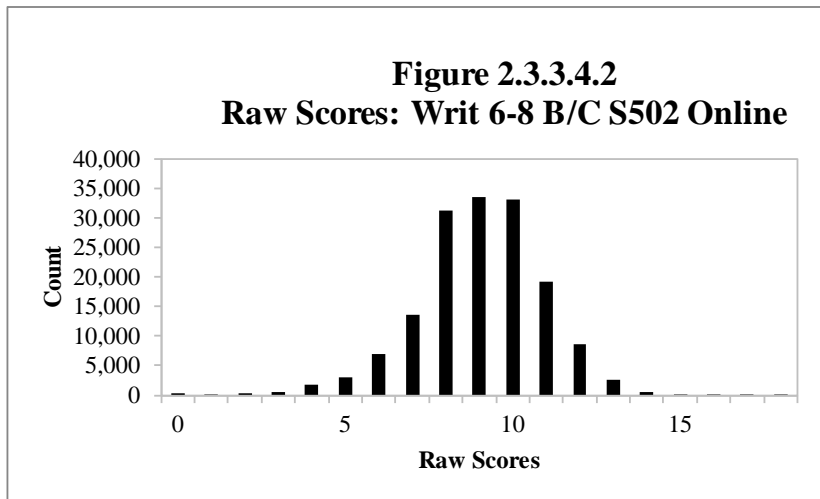


Table 2.3.3.4.2

Raw Score Descriptive Statistics: Writ 6-8 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	58,844	0	17	8.31	1.91
7	53,351	0	18	9.18	1.77
8	43,791	0	17	9.74	1.74
Total	155,986	0	18	9.01	1.91



2.3.3.5 Grades 9-12

Table 2.3.3.5.1

Raw Score Descriptive Statistics: Writ 9-12 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	21,579	0	16	6.71	3.28
10	18,973	0	15	6.79	3.07
11	13,751	0	17	7.60	2.79
12	9,077	0	16	7.90	2.72
Total	63,380	0	17	7.10	3.08

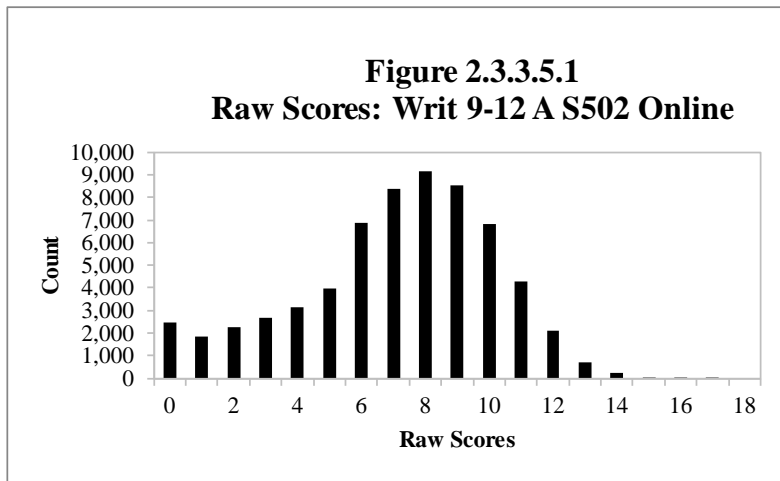
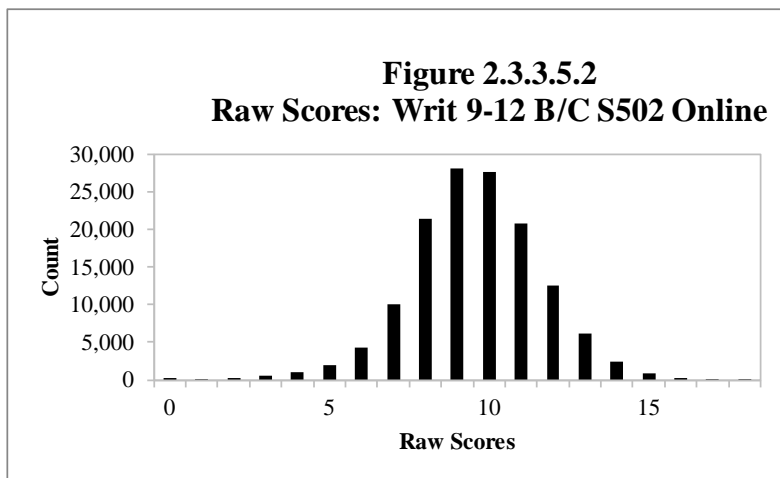


Table 2.3.3.5.2

Raw Score Descriptive Statistics: Writ 9-12 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	42,326	0	17	9.38	2.03
10	37,734	0	18	9.49	2.10
11	32,389	0	18	9.72	2.10
12	26,270	0	18	9.71	2.12
Total	138,719	0	18	9.55	2.09



2.3.4 Speaking

2.3.4.1 Grade 1

Table 2.3.4.1.1

Raw Score Descriptive Statistics: Spek 1 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	5,634	0	6	4.71	2.00
Total	5,634	0	6	4.71	2.00

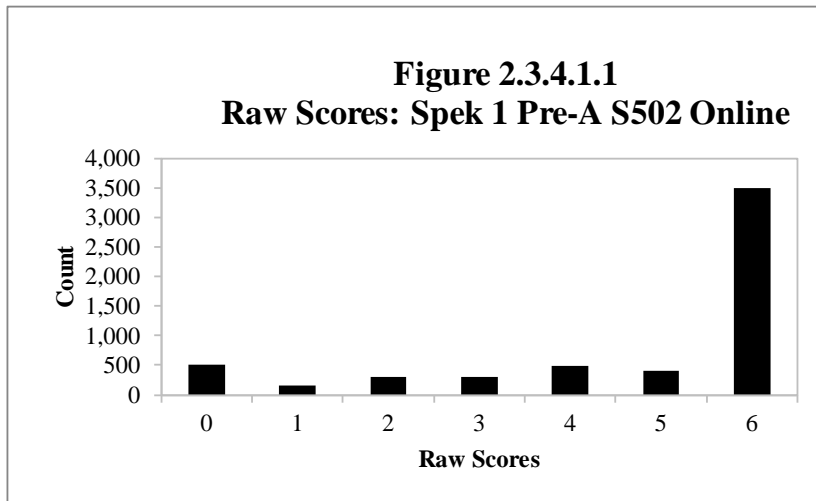


Table 2.3.4.1.2

Raw Score Descriptive Statistics: Spek 1 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	55,599	0	18	10.74	3.04
Total	55,599	0	18	10.74	3.04

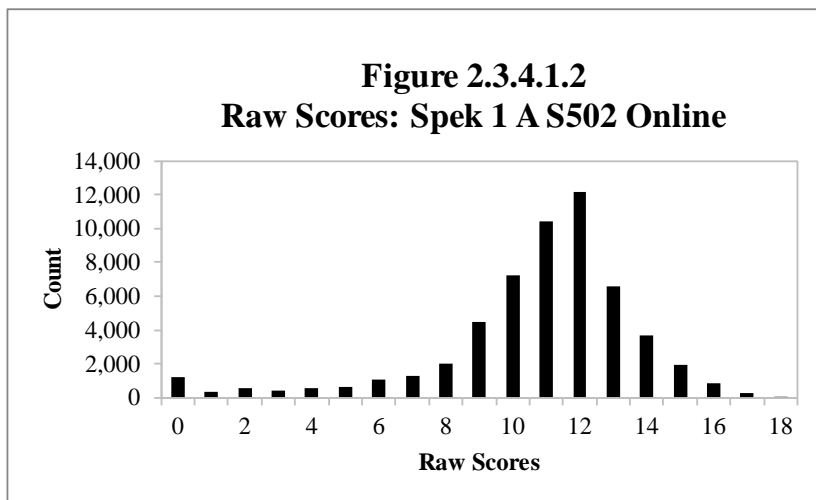
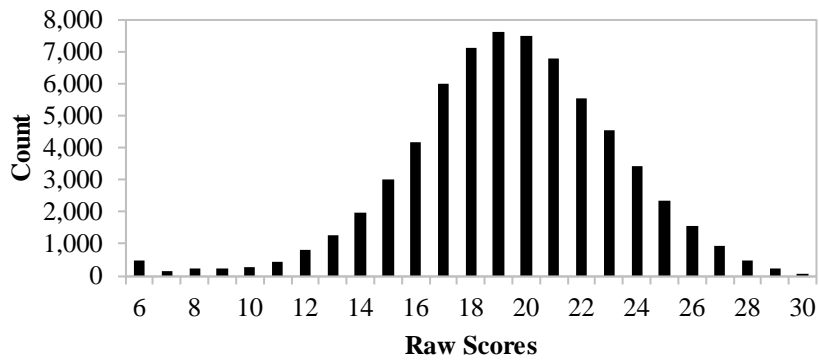


Table 2.3.4.1.3

Raw Score Descriptive Statistics: Spek 1 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	67,181	6	30	19.40	3.82
Total	67,181	6	30	19.40	3.82

Figure 2.3.4.1.3
Raw Scores: Spek 1 B/C S502 Online



2.3.4.2 Grades 2–3

Table 2.3.4.2.1

Raw Score Descriptive Statistics: Spek 2-3 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	3,050	0	6	5.16	1.67
3	6,101	0	6	5.23	1.62
Total	9,151	0	6	5.20	1.64

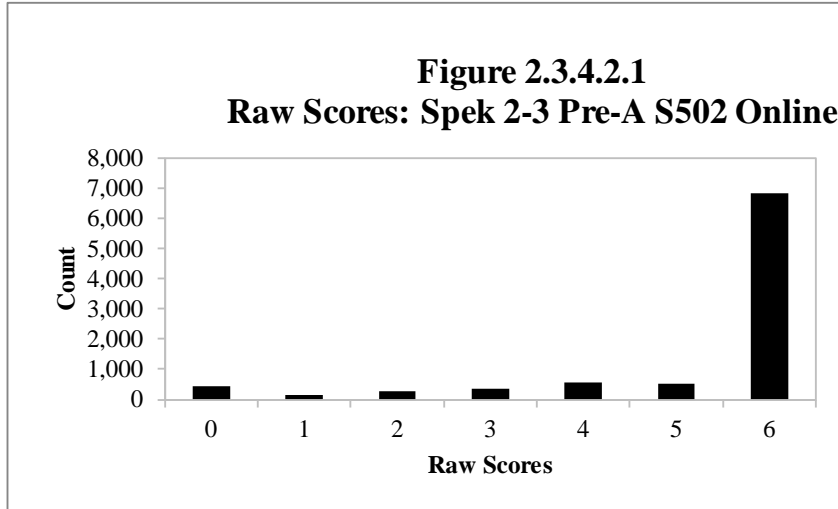


Table 2.3.4.2.2

Raw Score Descriptive Statistics: Spek 2-3 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	35,126	0	18	10.84	2.98
3	33,680	0	18	11.91	2.76
Total	68,806	0	18	11.36	2.92

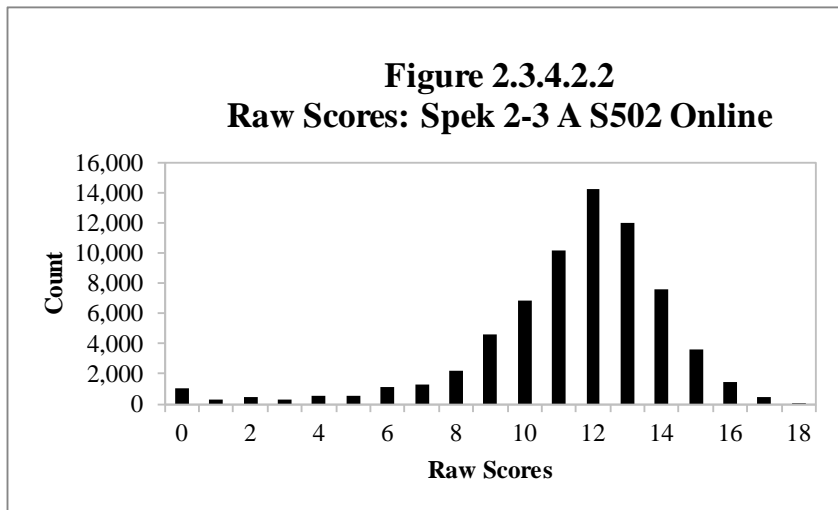
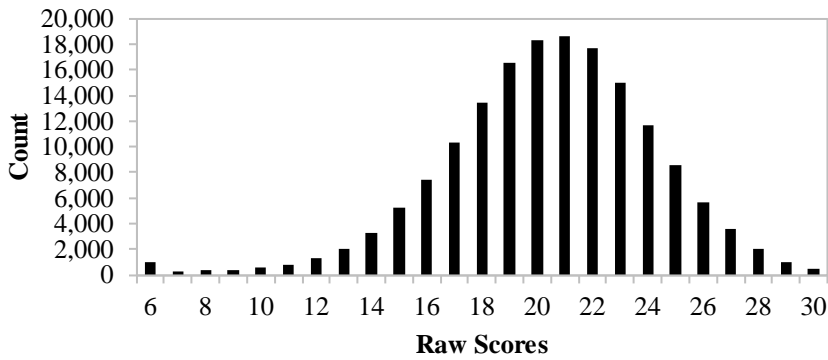


Table 2.3.4.2.3

Raw Score Descriptive Statistics: Spek 2-3 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	83,406	6	30	19.54	3.82
3	82,303	6	30	21.26	3.60
Total	165,709	6	30	20.40	3.81

Figure 2.3.4.2.3
Raw Scores: Spek 2-3 B/C S502 Online



2.3.4.3 Grades 4–5

Table 2.3.4.3.1

Raw Score Descriptive Statistics: Spek 4-5 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	1,247	0	6	4.43	2.08
5	1,990	0	6	4.66	1.94
Total	3,237	0	6	4.57	2.00

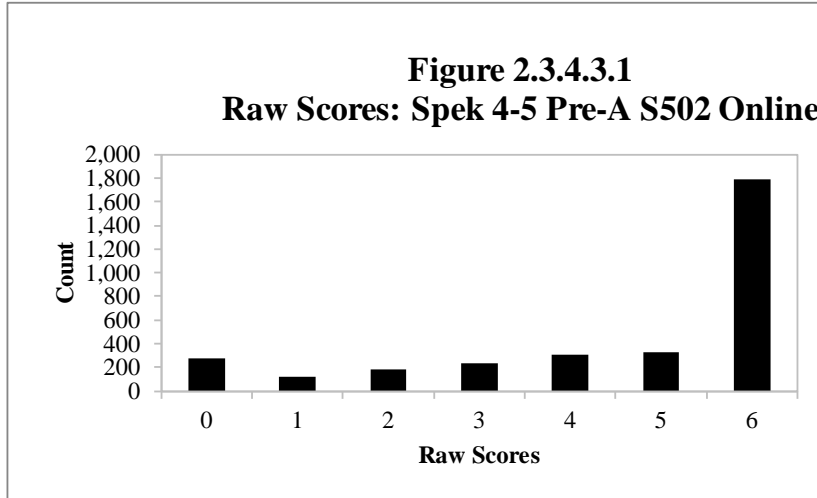


Table 2.3.4.3.2

Raw Score Descriptive Statistics: Spek 4-5 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	16,208	0	18	10.32	3.13
5	13,416	0	18	10.43	3.13
Total	29,624	0	18	10.37	3.13

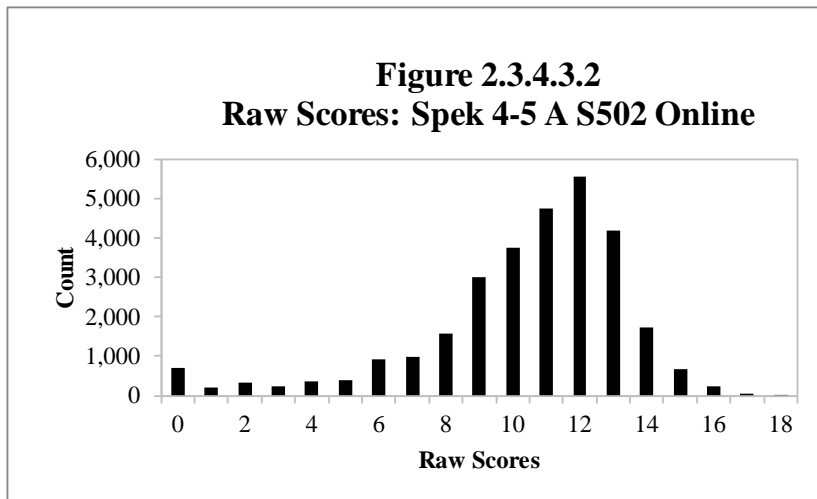
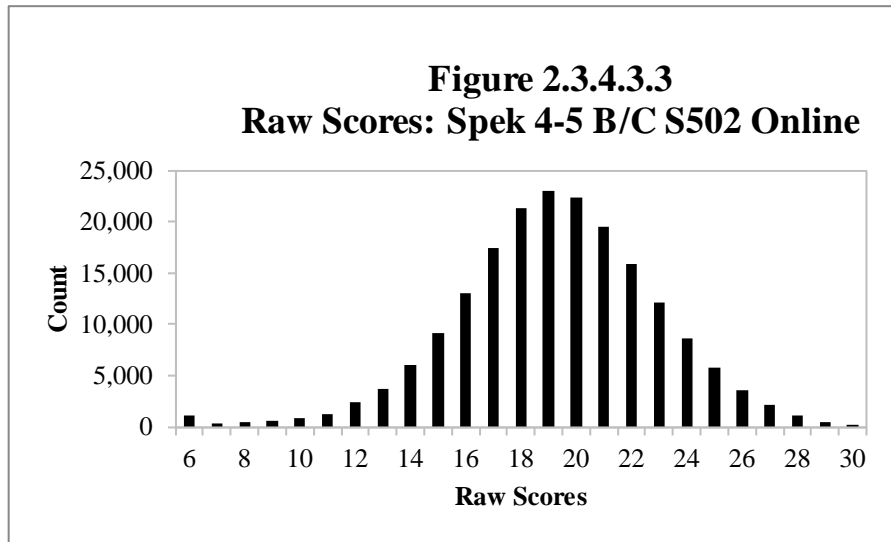


Table 2.3.4.3.3

Raw Score Descriptive Statistics: Spek 4-5 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	108,308	6	30	19.22	3.67
5	84,343	6	30	19.29	3.65
Total	192,651	6	30	19.25	3.66



2.3.4.4 Grades 6–8

Table 2.3.4.4.1

Raw Score Descriptive Statistics: Spek 6-8 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	1,275	0	6	4.88	1.89
7	1,661	0	6	4.74	1.97
8	2,828	0	6	4.94	1.84
Total	5,764	0	6	4.87	1.89

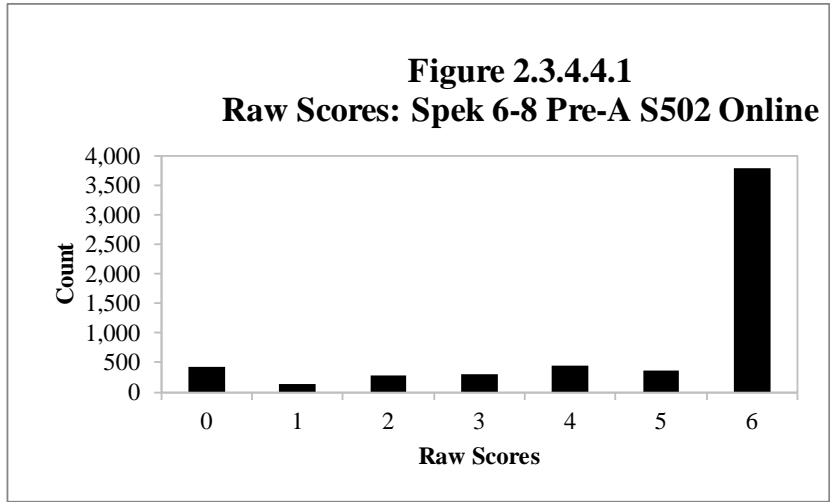


Table 2.3.4.4.2

Raw Score Descriptive Statistics: Spek 6-8 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	13,115	0	17	10.02	3.00
7	11,744	0	17	9.79	3.17
8	21,847	0	18	10.96	2.94
Total	46,706	0	18	10.40	3.06

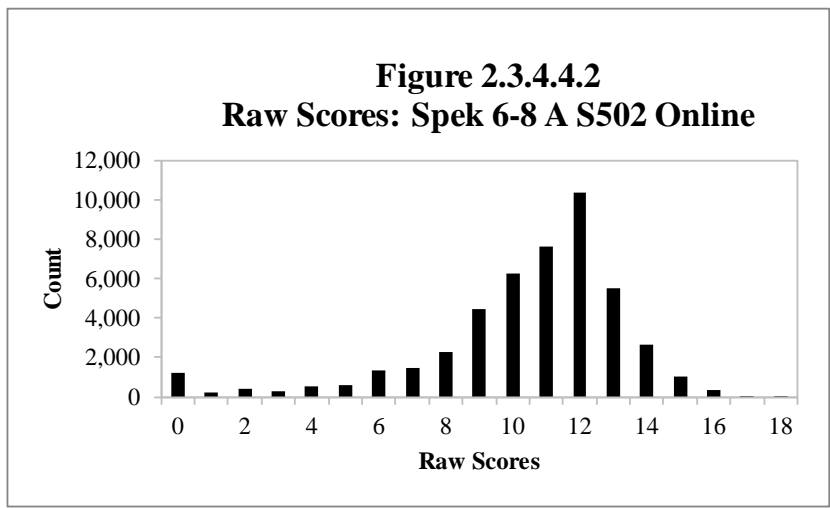
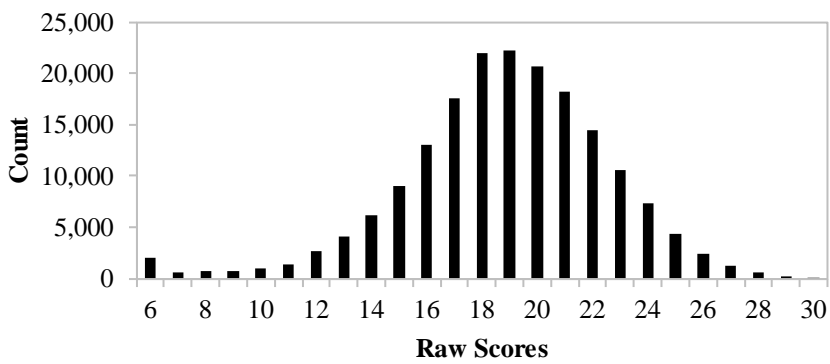


Table 2.3.4.4.3

Raw Score Descriptive Statistics: Spek 6-8 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	66,866	6	30	18.26	3.64
7	67,981	6	30	18.77	3.73
8	48,449	6	30	19.67	3.66
Total	183,296	6	30	18.82	3.72

Figure 2.3.4.4.3
Raw Scores: Spek 6-8 B/C S502 Online



2.3.4.5 Grades 9-12

Table 2.3.4.5.1

Raw Score Descriptive Statistics: Spek 9-12 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	2,366	0	6	4.74	1.91
10	3,951	0	6	5.16	1.64
11	3,125	0	6	5.41	1.44
12	2,546	0	6	5.47	1.44
Total	11,988	0	6	5.21	1.63

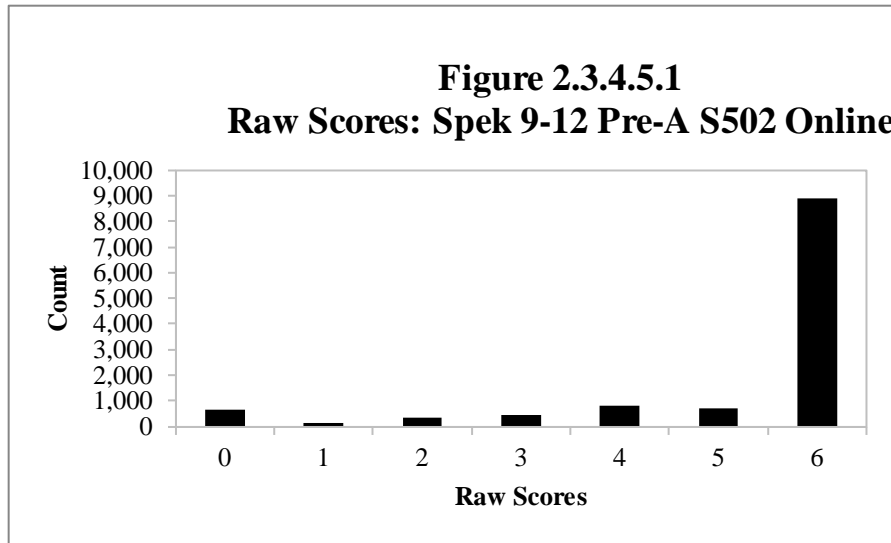


Table 2.3.4.5.2

Raw Score Descriptive Statistics: Spek 9-12 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	30,113	0	18	10.43	2.96
10	20,130	0	17	10.32	2.84
11	7,987	0	18	10.16	2.85
12	14,205	0	18	11.22	2.84
Total	72,435	0	18	10.52	2.91

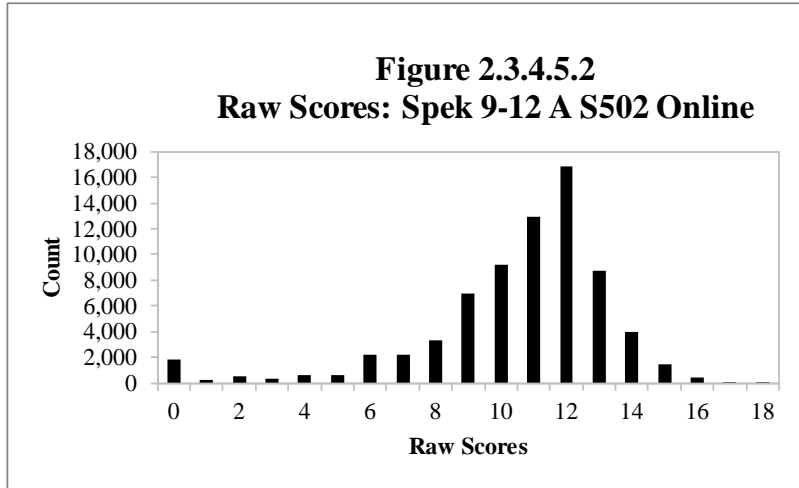
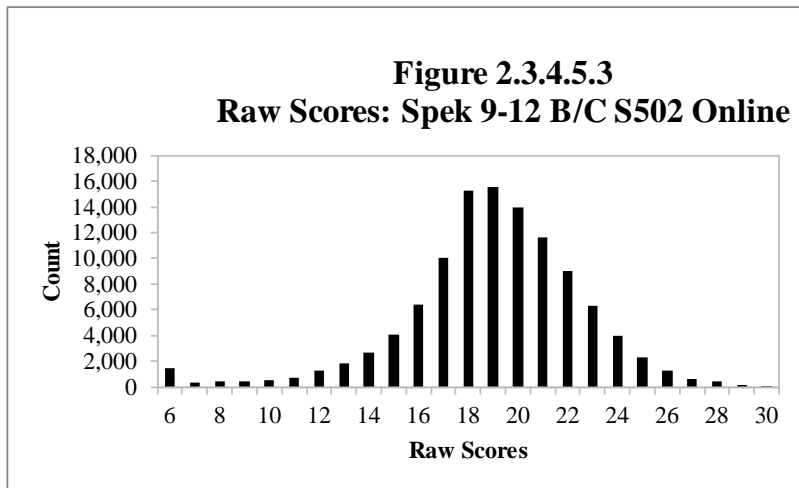


Table 2.3.4.5.3

Raw Score Descriptive Statistics: Spek 9-12 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	29,229	6	30	18.95	3.39
10	30,737	6	30	18.95	3.53
11	33,225	6	30	18.73	3.78
12	17,495	6	30	19.54	3.61
Total	110,686	6	30	18.97	3.59



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.).

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.

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 scores

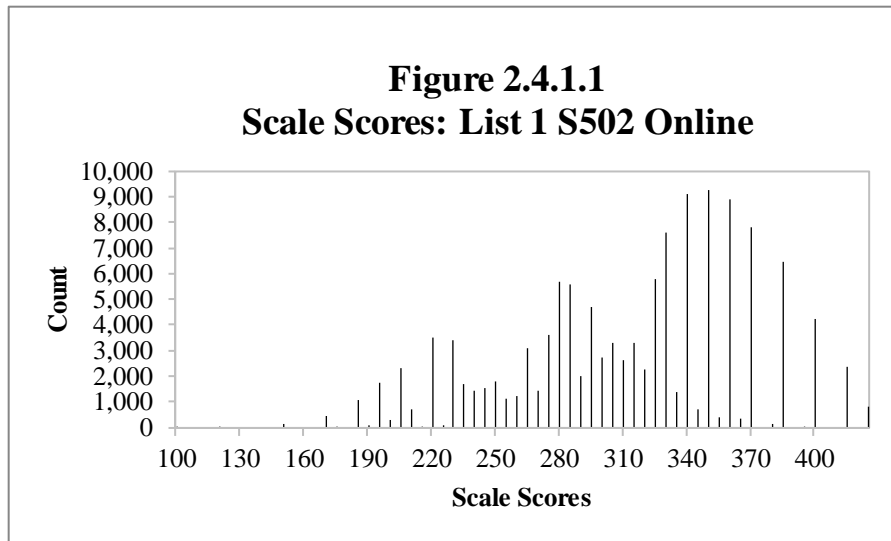
2.4.1 Listening

2.4.1.1 Grade 1

Table 2.4.1.1

Scale Score Descriptive Statistics: List 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	128,188	104	429	315.14	55.06
Total	128,188	104	429	315.14	55.06

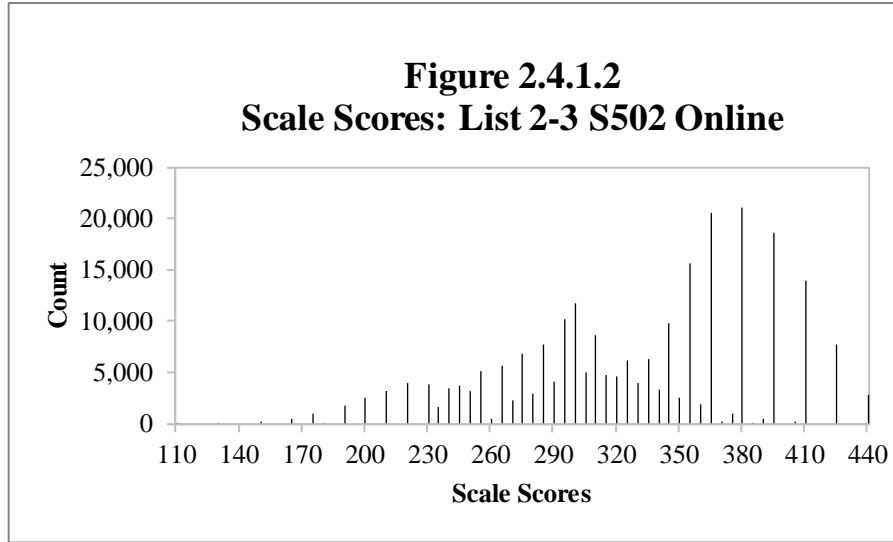


2.4.1.2 Grades 2–3

Table 2.4.1.2

Scale Score Descriptive Statistics: List 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	122,368	112	442	319.03	57.36
3	122,284	112	442	343.21	58.43
Total	244,652	112	442	331.11	59.15

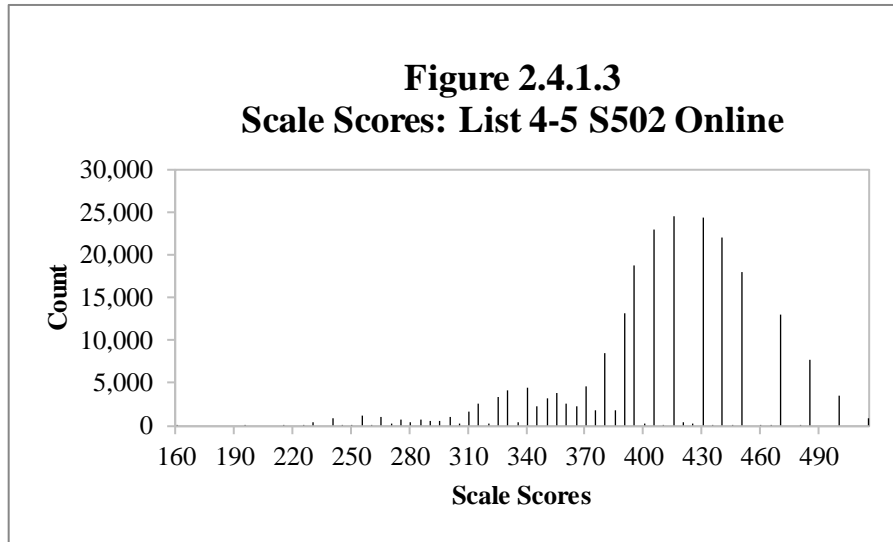


2.4.1.3 Grades 4–5

Table 2.4.1.3

Scale Score Descriptive Statistics: List 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	126,013	164	518	406.99	46.96
5	99,811	164	518	411.31	50.69
Total	225,824	164	518	408.90	48.69

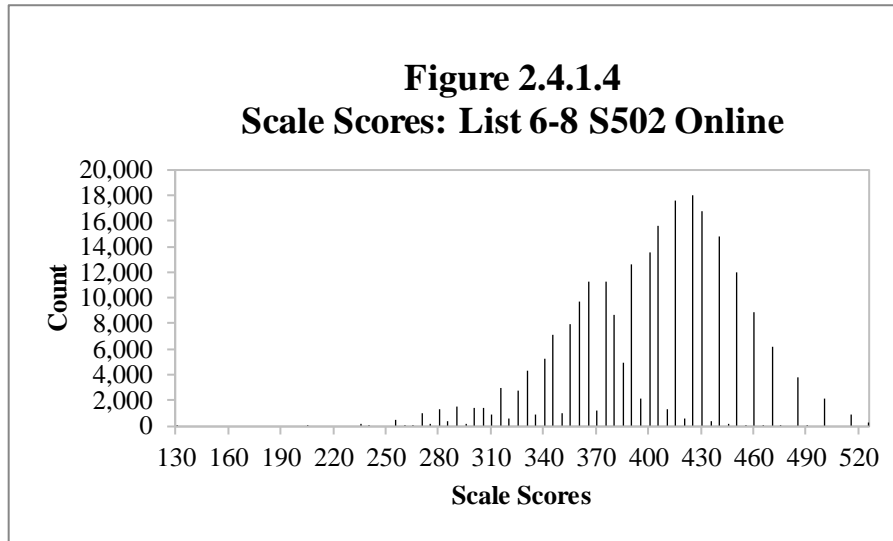


2.4.1.4 Grades 6–8

Table 2.4.1.4

Scale Score Descriptive Statistics: List 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	81,591	132	529	395.75	41.17
7	81,572	132	529	401.23	45.89
8	73,325	132	529	406.01	49.80
Total	236,488	132	529	400.82	45.80

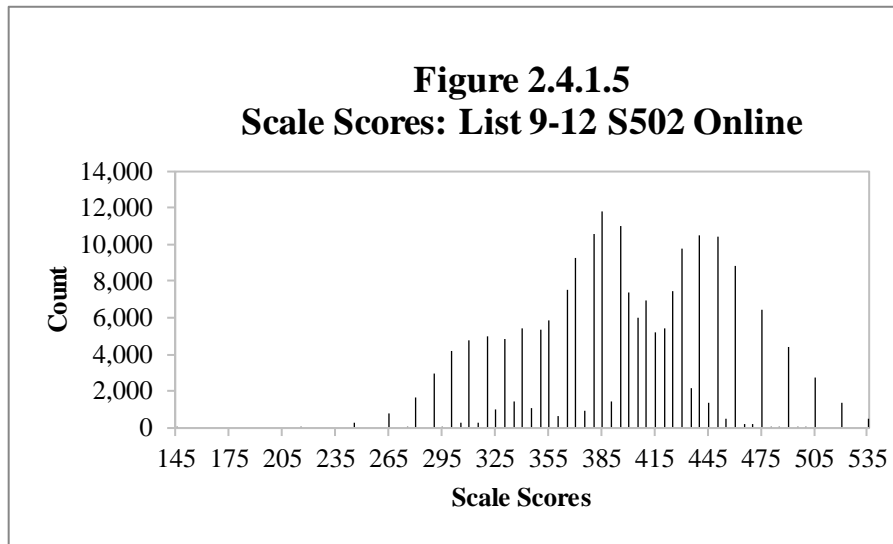


2.4.1.5 Grades 9–12

Table 2.4.1.5

Scale Score Descriptive Statistics: List 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	61,624	148	535	395.17	50.59
10	55,044	218	535	396.48	52.99
11	44,872	218	535	403.30	51.78
12	34,341	218	535	404.63	52.30
Total	195,881	148	535	399.06	52.00



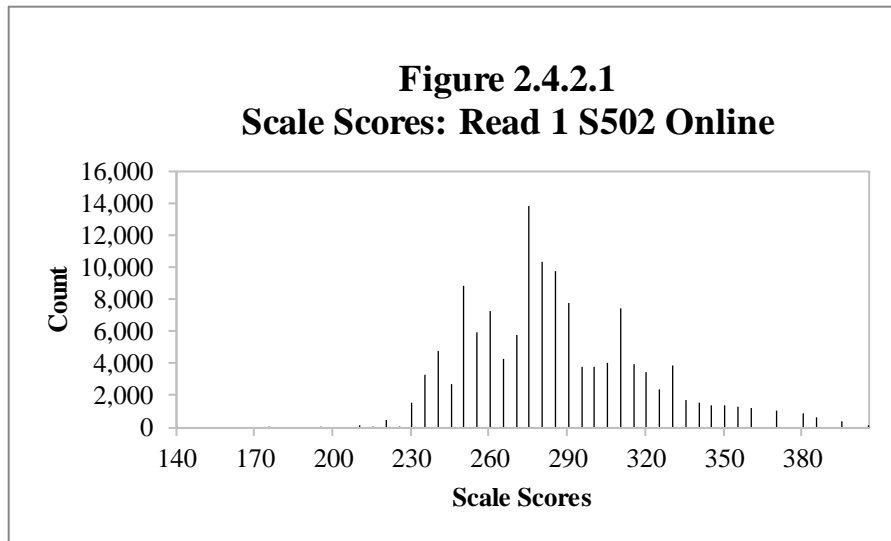
2.4.2 Reading

2.4.2.1 Grade 1

Table 2.4.2.1

Scale Score Descriptive Statistics: Read 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	131,078	141	407	287.65	33.27
Total	131,078	141	407	287.65	33.27

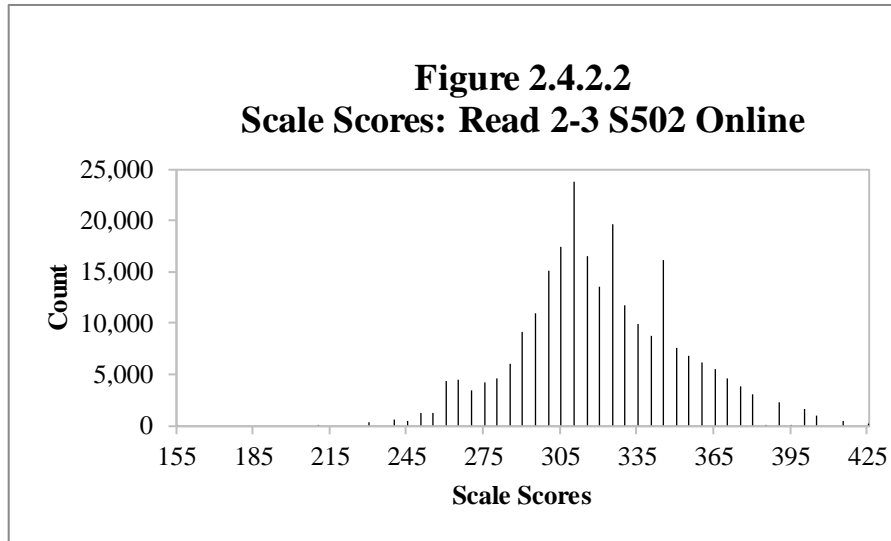


2.4.2.2 Grades 2–3

Table 2.4.2.2

Scale Score Descriptive Statistics: Read 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	123,835	158	427	316.88	27.25
3	123,099	158	427	326.68	32.78
Total	246,934	158	427	321.77	30.53

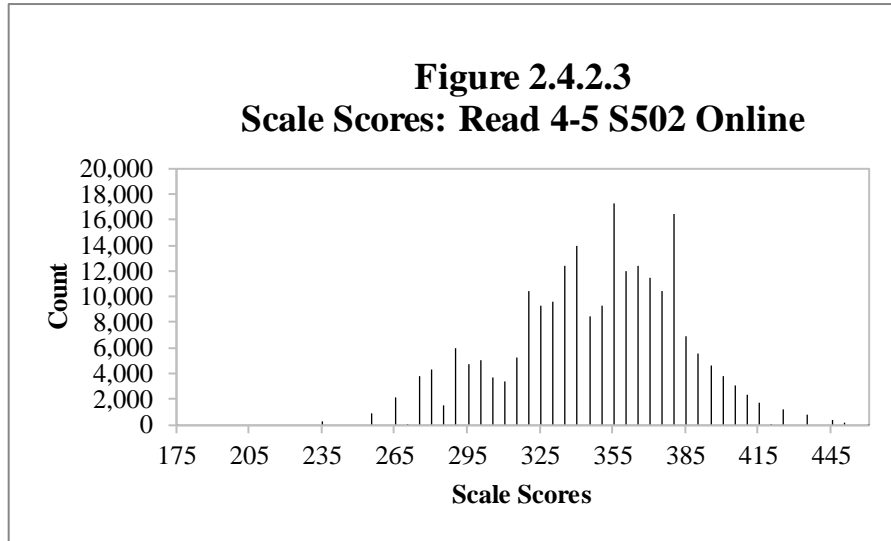


2.4.2.3 Grades 4–5

Table 2.4.2.3

Scale Score Descriptive Statistics: Read 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	125,652	175	463	347.72	34.53
5	99,514	175	463	351.97	35.72
Total	225,166	175	463	349.60	35.13

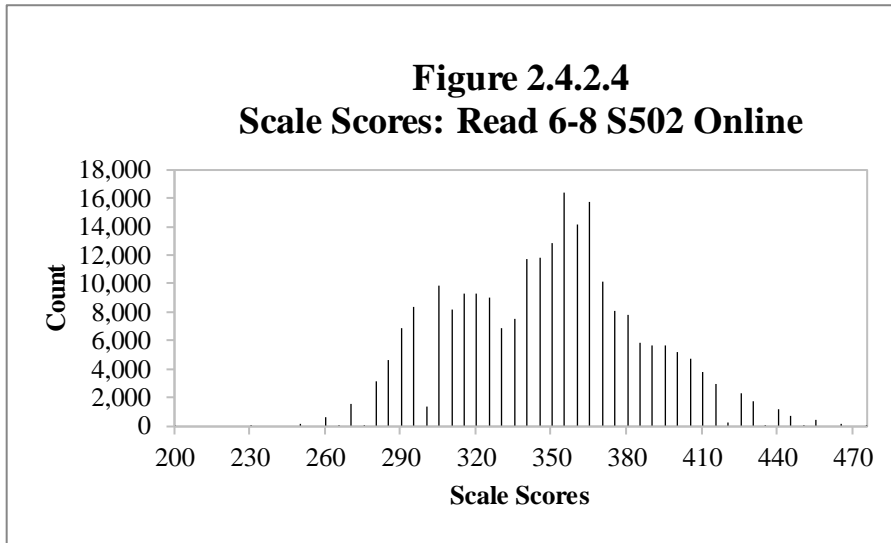


2.4.2.4 Grades 6–8

Table 2.4.2.4

Scale Score Descriptive Statistics: Read 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	81,786	200	476	342.61	33.29
7	81,424	200	476	351.99	36.89
8	73,049	200	476	357.34	39.95
Total	236,259	200	476	350.40	37.19

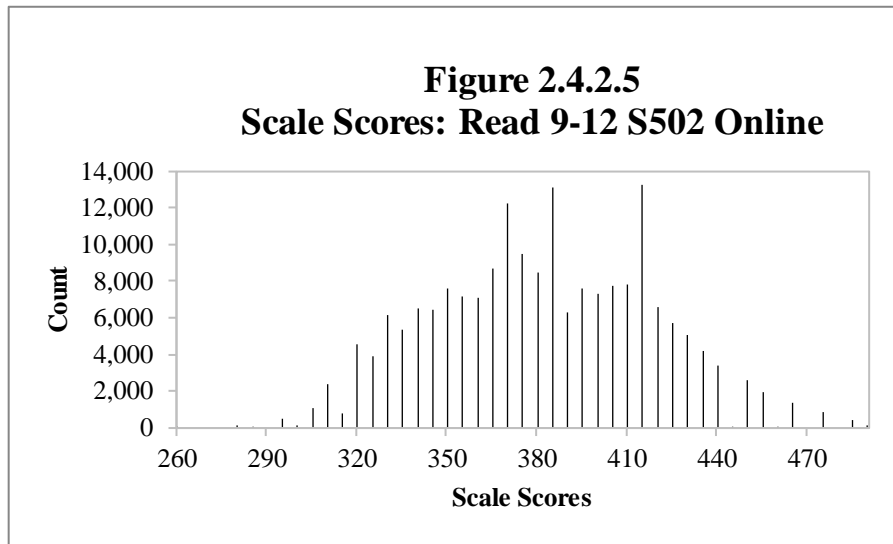


2.4.2.5 Grades 9–12

Table 2.4.2.5

Scale Score Descriptive Statistics: Read 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	61,038	261	494	379.06	35.43
10	54,455	261	494	382.01	36.97
11	44,266	261	494	387.62	36.90
12	34,009	261	494	389.83	36.52
Total	193,768	261	494	383.73	36.64



2.4.3 Writing

2.4.3.1 Grade 1

Table 2.4.3.1.1

Scale Score Descriptive Statistics: Writ 1 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	115,655	111	338	235.46	37.55
Total	115,655	111	338	235.46	37.55

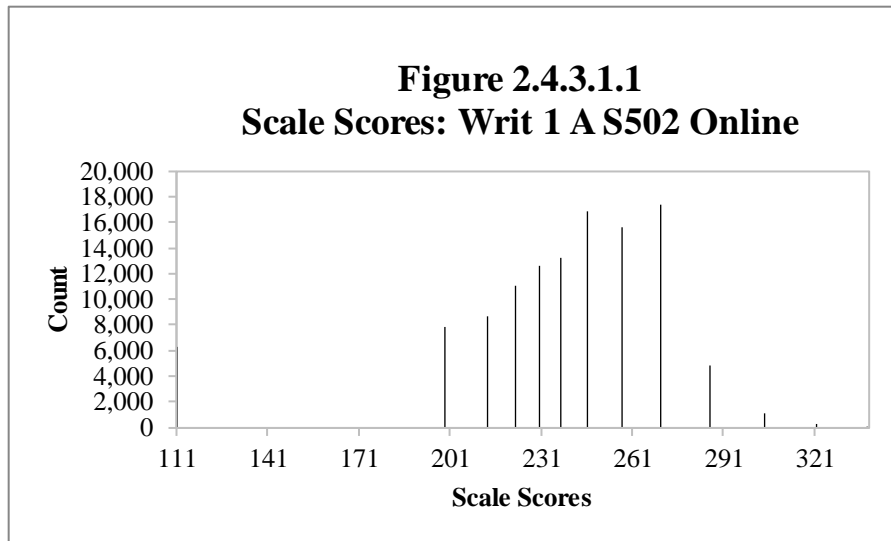


Table 2.4.3.1.2

Scale Score Descriptive Statistics: Writ 1 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	21,518	111	428	287.86	33.55
Total	21,518	111	428	287.86	33.55

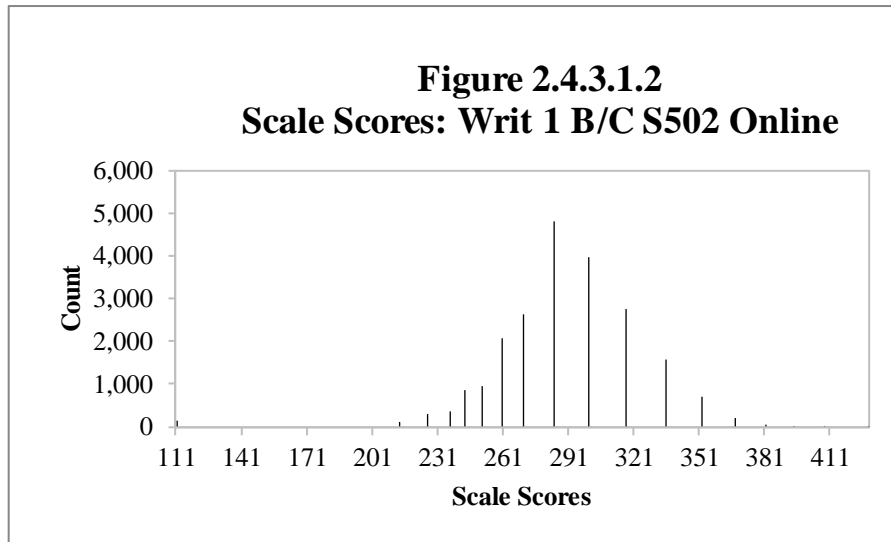
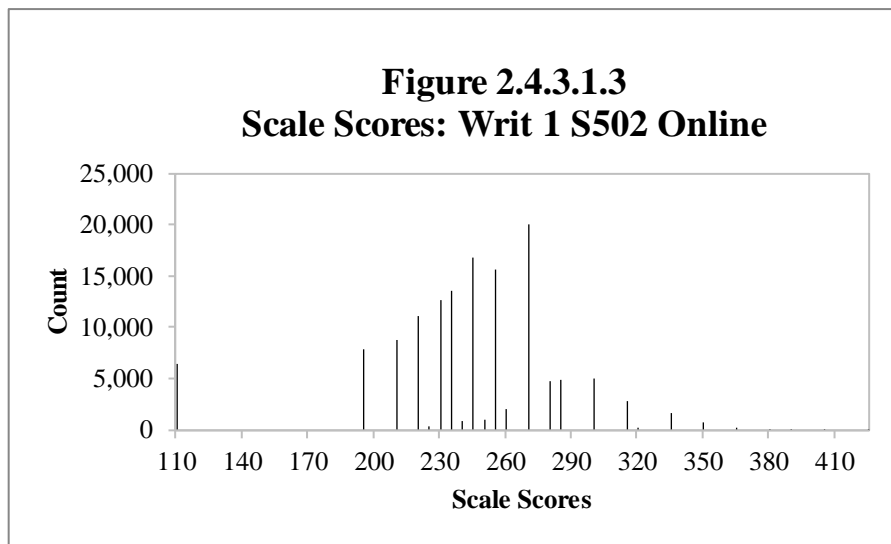


Table 2.4.3.1.3

Scale Score Descriptive Statistics: Writ 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	137,173	111	428	243.68	41.58
Total	137,173	111	428	243.68	41.58



2.4.3.2 Grades 2–3

Table 2.4.3.2.1

Scale Score Descriptive Statistics: Writ 2-3 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	41,920	133	383	249.84	48.13
3	31,972	133	383	263.75	47.94
Total	73,892	133	383	255.86	48.54

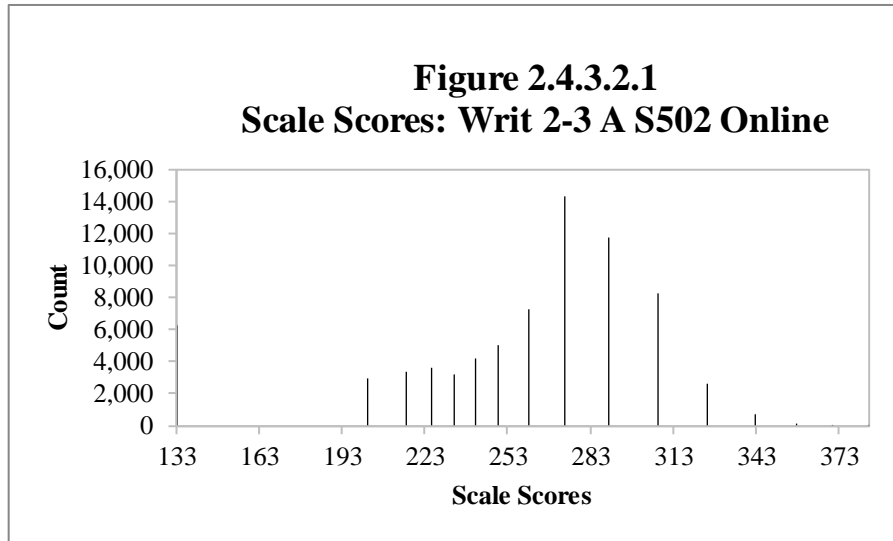


Table 2.4.3.2.2

Scale Score Descriptive Statistics: Writ 2-3 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	89,098	133	422	301.58	36.43
3	98,239	133	453	322.79	28.32
Total	187,337	133	453	312.71	34.12

Figure 2.4.3.2.2
Scale Scores: Writ 2-3 B/C S502 Online

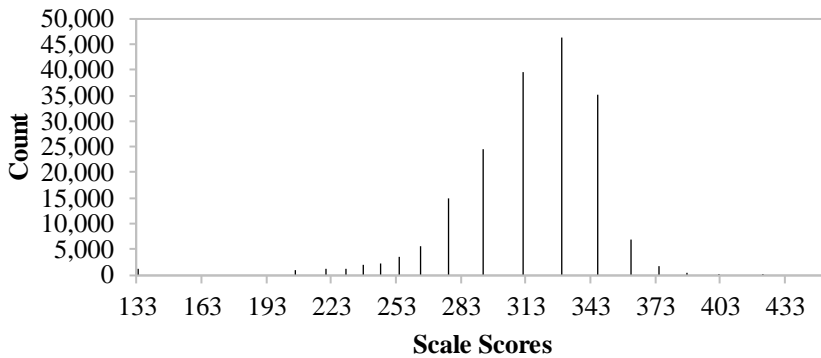
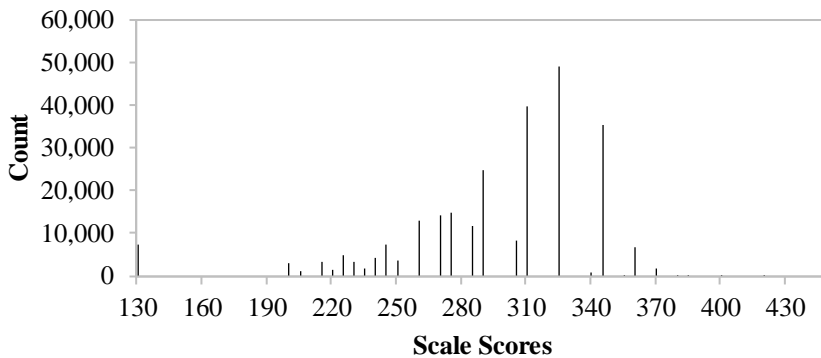


Table 2.4.3.2.3

Scale Score Descriptive Statistics: Writ 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	131,018	133	422	285.03	47.18
3	130,211	133	453	308.30	42.60
Total	261,229	133	453	296.63	46.44

Figure 2.4.3.2.3
Scale Scores: Writ 2-3 S502 Online



2.4.3.3 Grades 4–5

Table 2.4.3.3.1

Scale Score Descriptive Statistics: Writ 4-5 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	23,547	155	416	271.36	51.18
5	22,475	155	416	283.21	49.17
Total	46,022	155	416	277.15	50.56

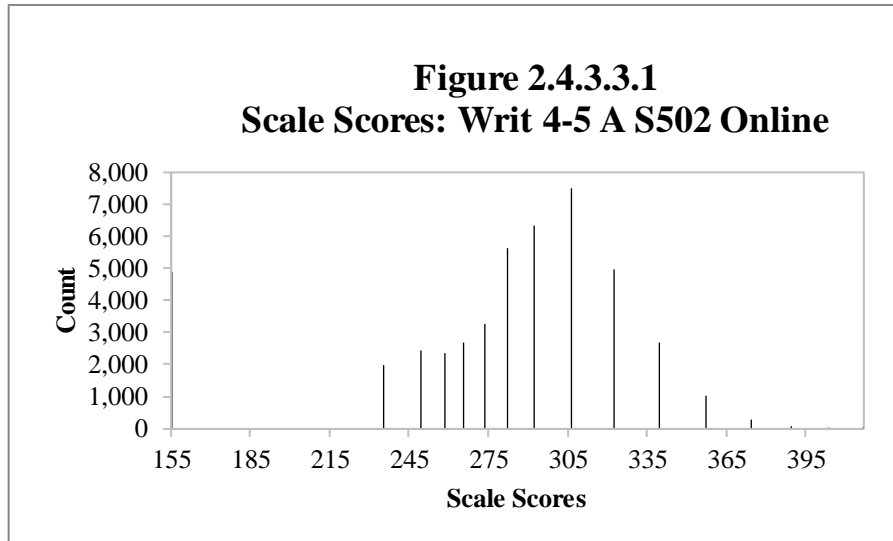


Table 2.4.3.3.2

Scale Score Descriptive Statistics: Writ 4-5 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	107,178	155	488	342.64	32.62
5	80,865	155	488	354.24	30.10
Total	188,043	155	488	347.63	32.08

Figure 2.4.3.3.2
Scale Scores: Writ 4-5 B/C S502 Online

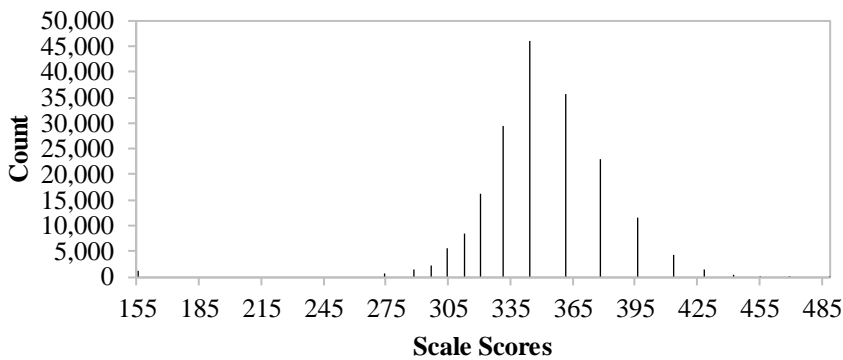
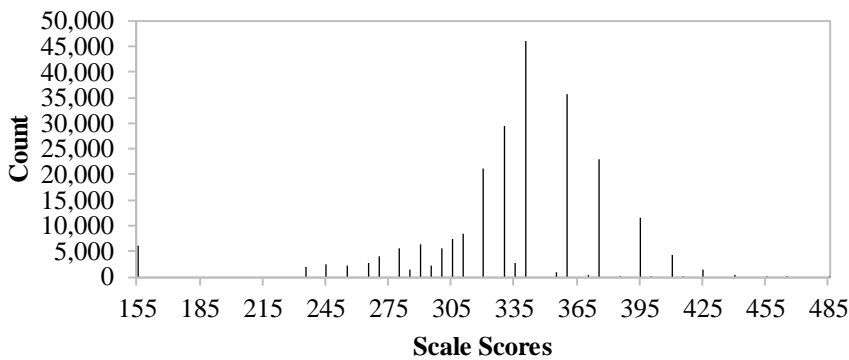


Table 2.4.3.3.3

Scale Score Descriptive Statistics: Writ 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	130,725	155	488	329.80	45.76
5	103,340	155	488	338.79	45.75
Total	234,065	155	488	333.77	45.98

Figure 2.4.3.3.3
Scale Scores: Writ 4-5 S502 Online



2.4.3.4 Grades 6–8

Table 2.4.3.4.1

Scale Score Descriptive Statistics: Writ 6-8 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	26,020	188	411	287.96	35.43
7	31,064	188	411	296.85	36.24
8	31,503	188	425	302.95	36.89
Total	88,587	188	425	296.41	36.74

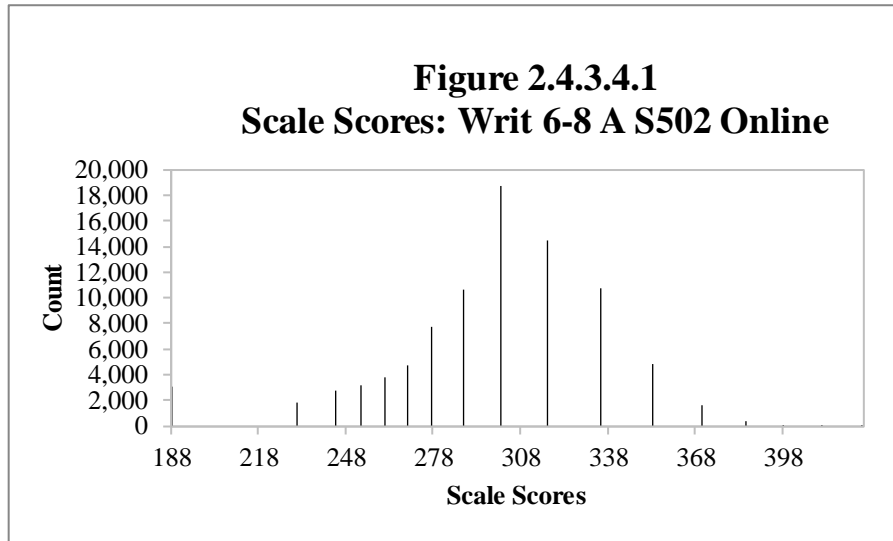


Table 2.4.3.4.2

Scale Score Descriptive Statistics: Writ 6-8 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	58,844	188	461	324.79	27.17
7	53,351	188	492	337.74	27.30
8	43,791	188	461	346.61	27.73
Total	155,986	188	492	335.34	28.79

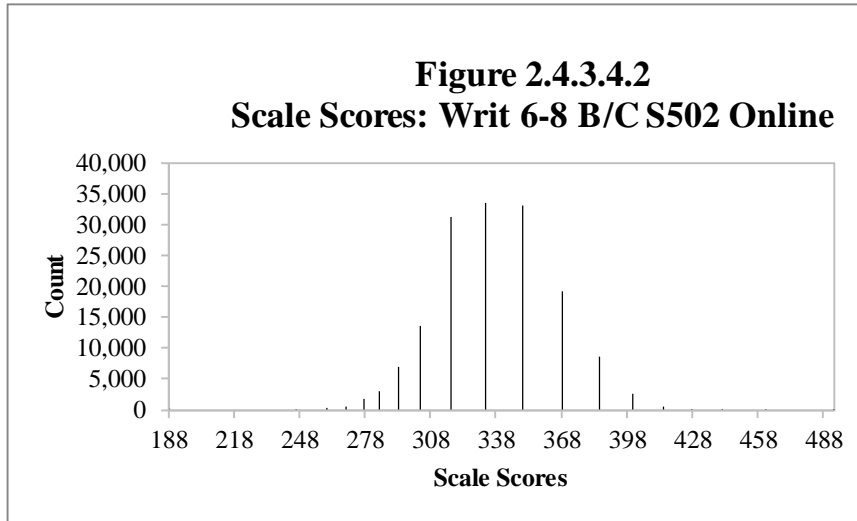
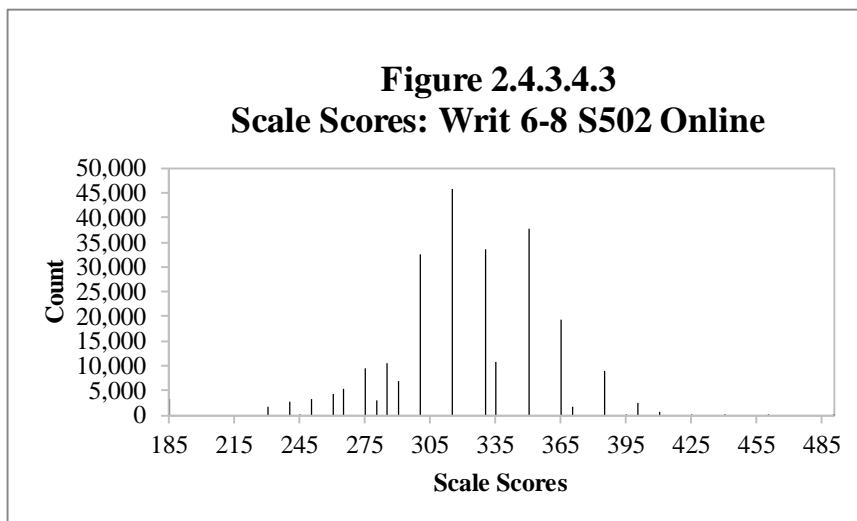


Table 2.4.3.4.3

Scale Score Descriptive Statistics: Writ 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	84,864	188	461	313.50	34.43
7	84,415	188	492	322.69	36.65
8	75,294	188	461	328.34	38.48
Total	244,573	188	492	321.24	36.98



2.4.3.5 Grades 9–12

Table 2.4.3.5.1

Scale Score Descriptive Statistics: Writ 9-12 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	21,579	232	456	320.97	41.38
10	18,973	232	441	321.71	38.76
11	13,751	232	476	331.54	36.67
12	9,077	232	456	335.39	36.58
Total	63,380	232	476	325.55	39.36

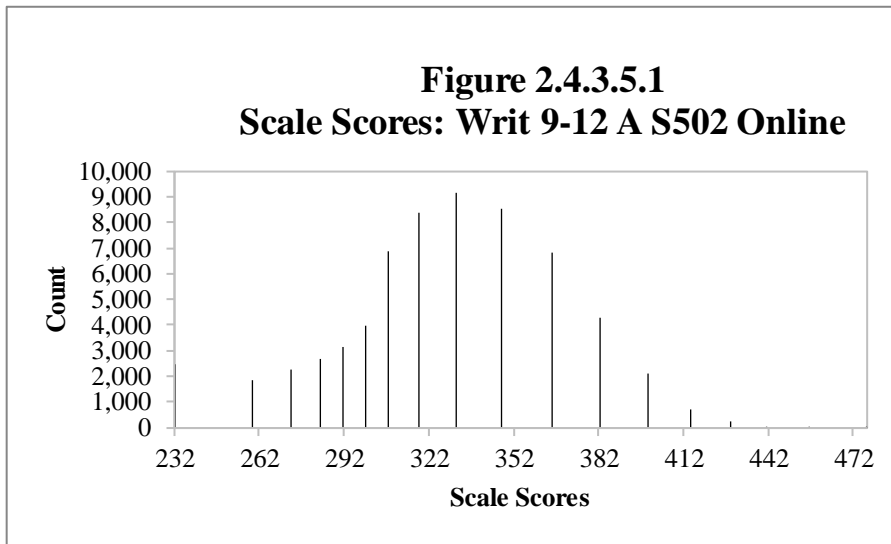


Table 2.4.3.5.2

Scale Score Descriptive Statistics: Writ 9-12 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	42,326	232	475	355.13	30.90
10	37,734	232	507	357.01	32.00
11	32,389	232	507	360.56	32.32
12	26,270	232	507	360.45	32.54
Total	138,719	232	507	357.92	31.93

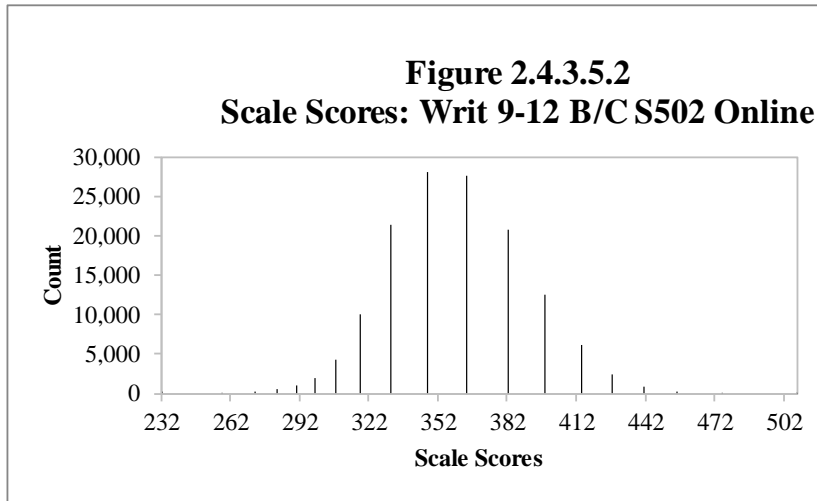
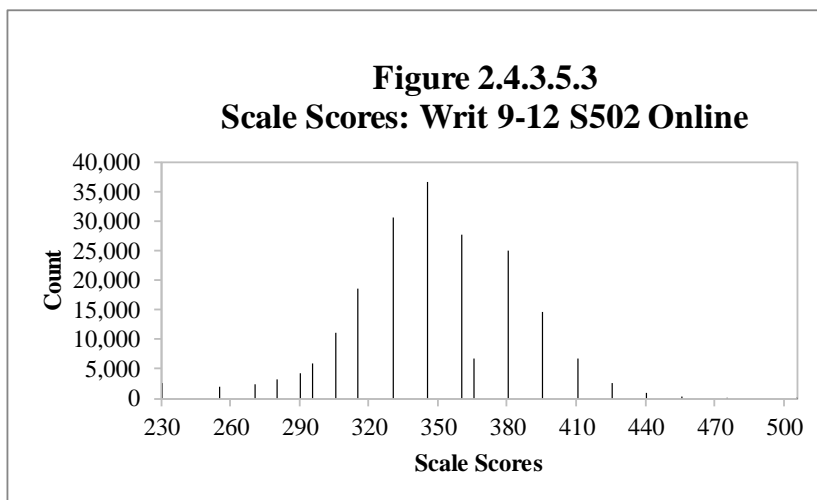


Table 2.4.3.5.3

Scale Score Descriptive Statistics: Writ 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	63,905	232	475	343.59	38.36
10	56,707	232	507	345.20	38.23
11	46,140	232	507	351.91	36.19
12	35,347	232	507	354.02	35.36
Total	202,099	232	507	347.77	37.57



2.4.4 Speaking

2.4.4.1 Grade 1

Table 2.4.4.1.1

Scale Score Descriptive Statistics: Spek 1 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	5,634	106	176	159.79	24.63
Total	5,634	106	176	159.79	24.63

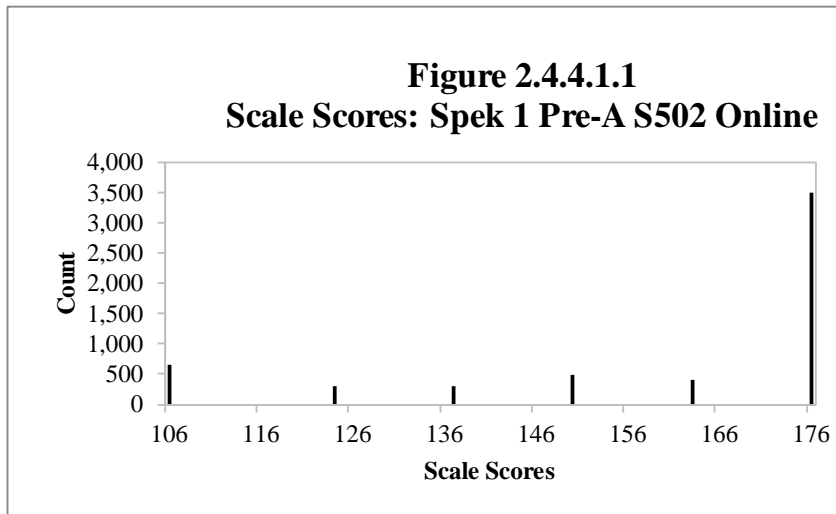


Table 2.4.4.1.2

Scale Score Descriptive Statistics: Spek 1 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	55,599	106	380	230.35	48.93
Total	55,599	106	380	230.35	48.93

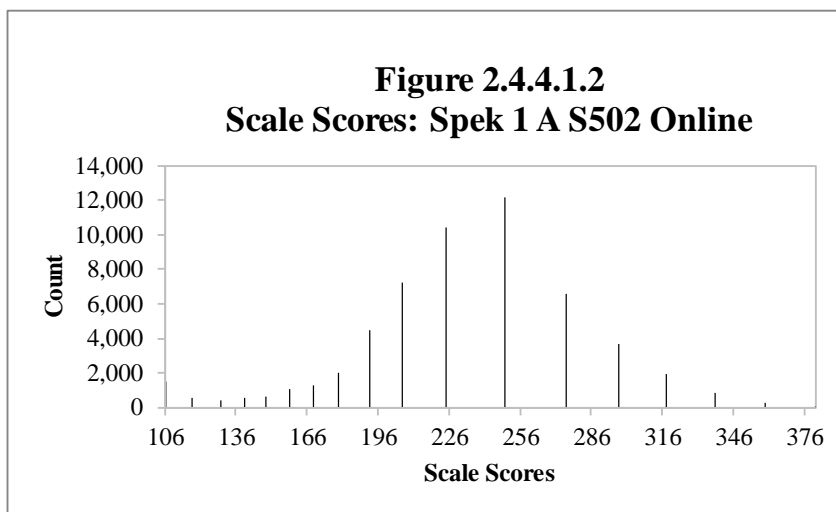


Table 2.4.4.1.3

Scale Score Descriptive Statistics: Spek 1 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	67,181	106	403	270.68	43.24
Total	67,181	106	403	270.68	43.24

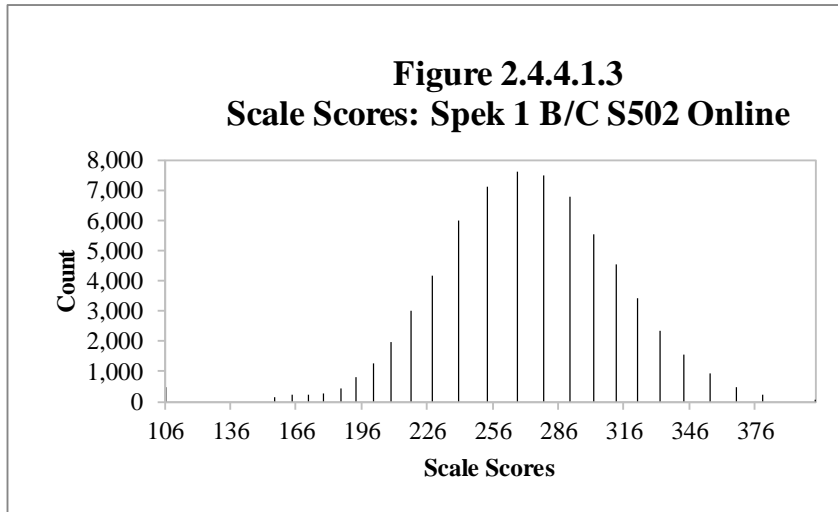
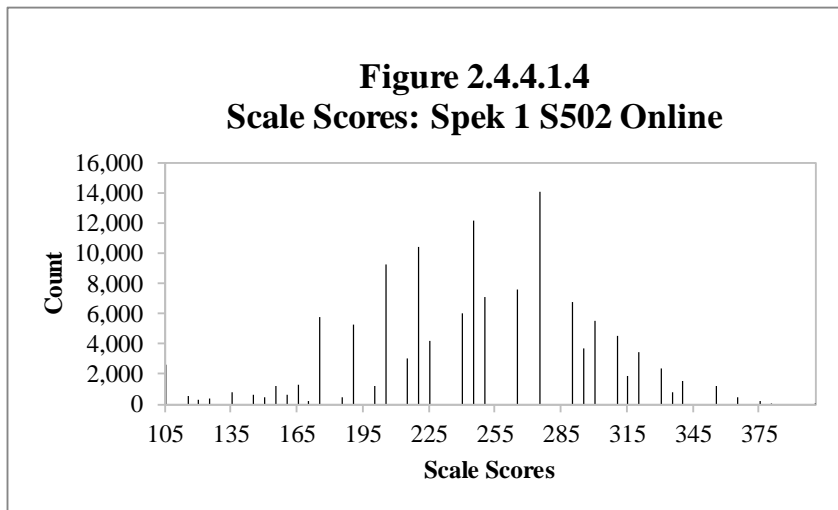


Table 2.4.4.1.4

Scale Score Descriptive Statistics: Spek 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	128,414	106	403	248.36	52.79
Total	128,414	106	403	248.36	52.79



2.4.4.2 Grades 2–3

Table 2.4.4.2.1

Scale Score Descriptive Statistics: Spek 2-3 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	3,050	118	172	162.13	18.30
3	6,101	118	172	162.90	17.75
Total	9,151	118	172	162.64	17.94

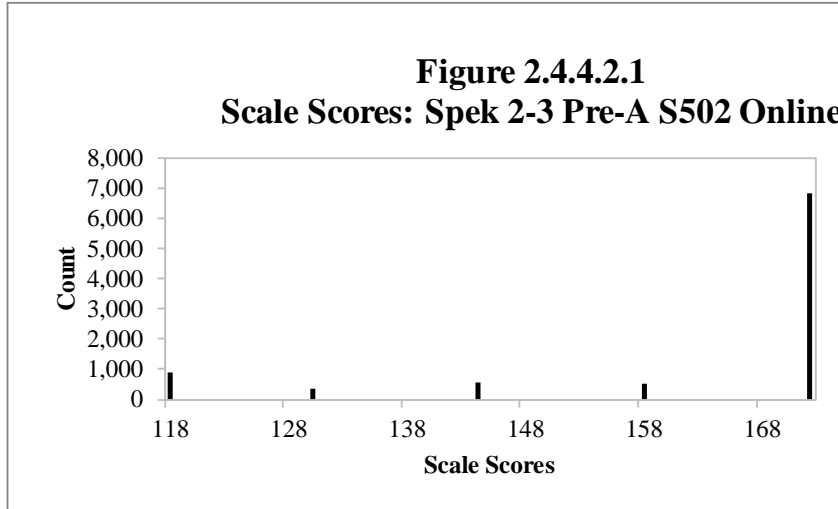


Table 2.4.4.2.2

Scale Score Descriptive Statistics: Spek 2-3 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	35,126	118	378	232.50	48.90
3	33,680	118	378	253.23	49.10
Total	68,806	118	378	242.65	50.08

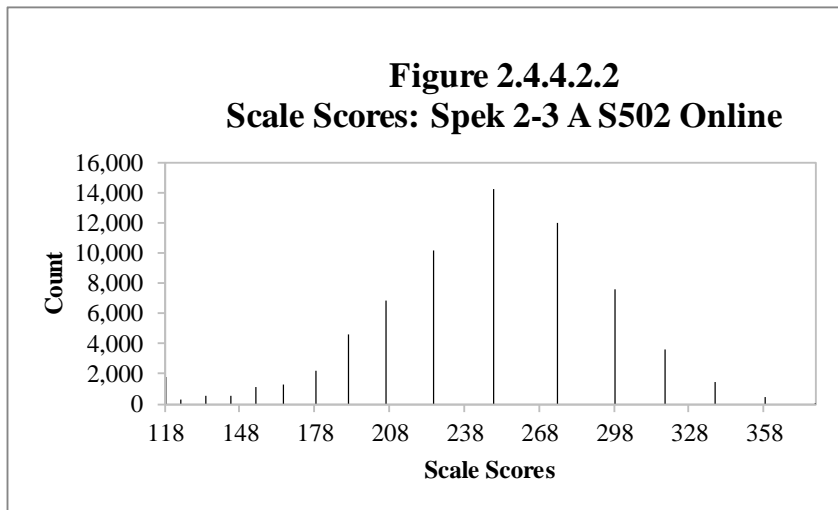


Table 2.4.4.2.3

Scale Score Descriptive Statistics: Spek 2-3 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	83,406	118	425	277.06	43.33
3	82,303	118	425	296.72	41.05
Total	165,709	118	425	286.82	43.34

Figure 2.4.4.2.3
Scale Scores: Spek 2-3 B/C S502 Online

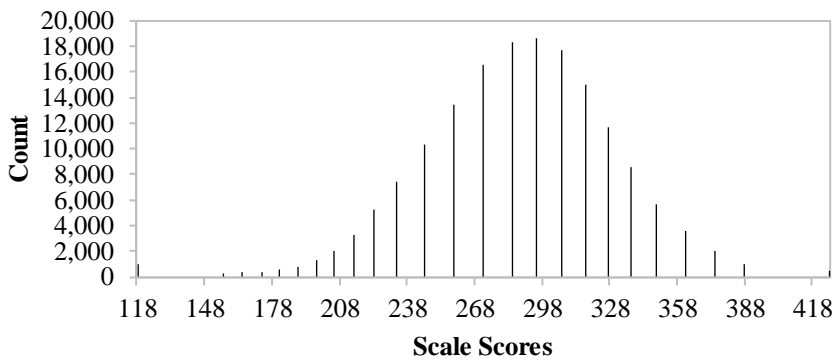
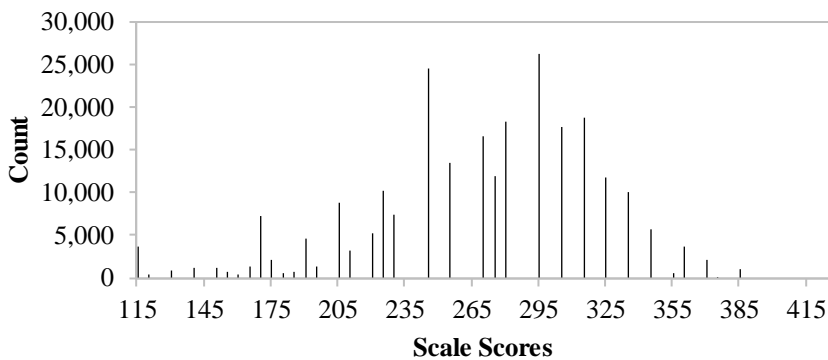


Table 2.4.4.2.4

Scale Score Descriptive Statistics: Spek 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	121,582	118	425	261.30	51.42
3	122,084	118	425	278.03	53.71
Total	243,666	118	425	269.68	53.24

Figure 2.4.4.2.4
Scale Scores: Spek 2-3 S502 Online



2.4.4.3 Grades 4–5

Table 2.4.4.3.1

Scale Score Descriptive Statistics: Spek 4-5 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	1,247	130	203	182.84	26.36
5	1,990	130	203	185.86	24.65
Total	3,237	130	203	184.70	25.36

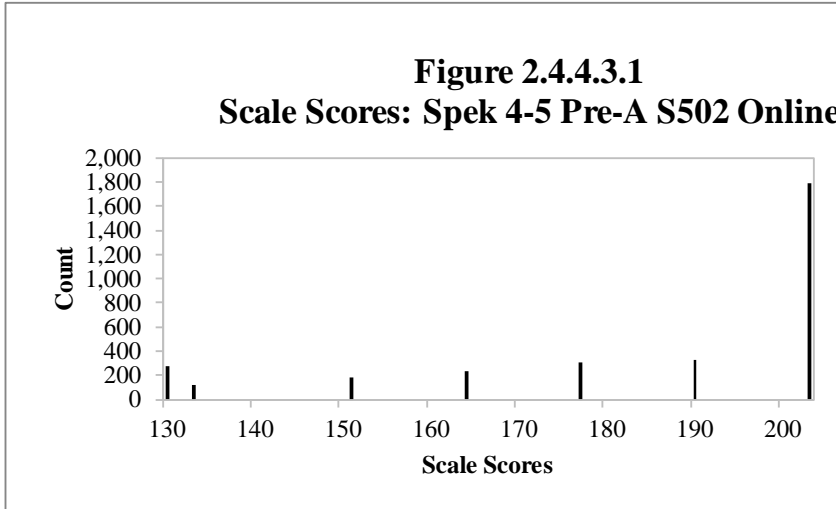


Table 2.4.4.3.2

Scale Score Descriptive Statistics: Spek 4-5 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	16,208	130	415	257.20	50.70
5	13,416	130	415	259.36	50.95
Total	29,624	130	415	258.18	50.82

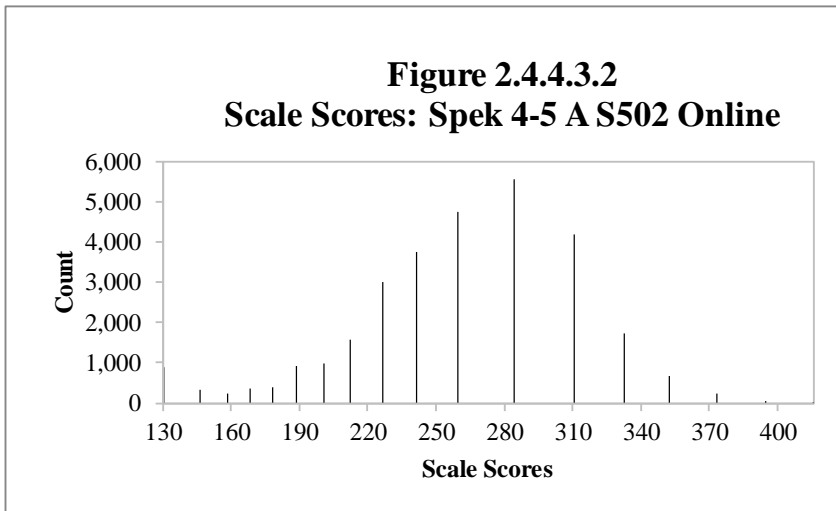


Table 2.4.4.3.3

Scale Score Descriptive Statistics: Spek 4-5 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	108,308	130	443	311.98	42.25
5	84,343	130	443	312.69	42.17
Total	192,651	130	443	312.29	42.22

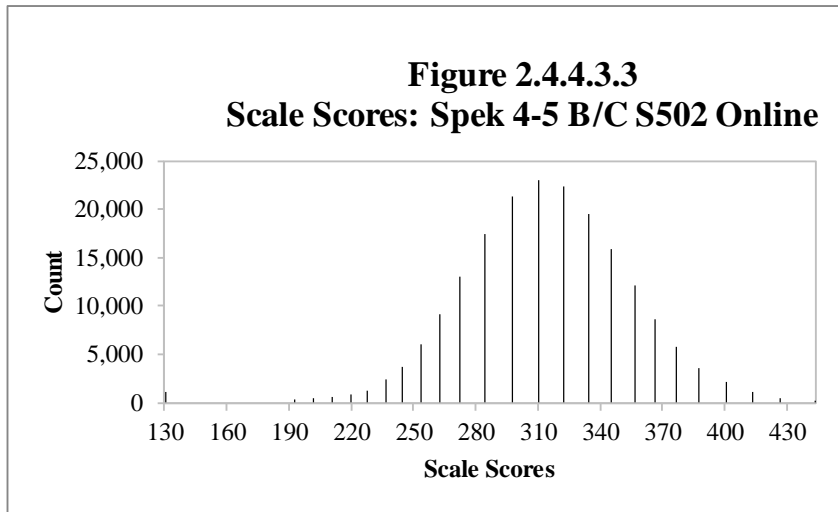
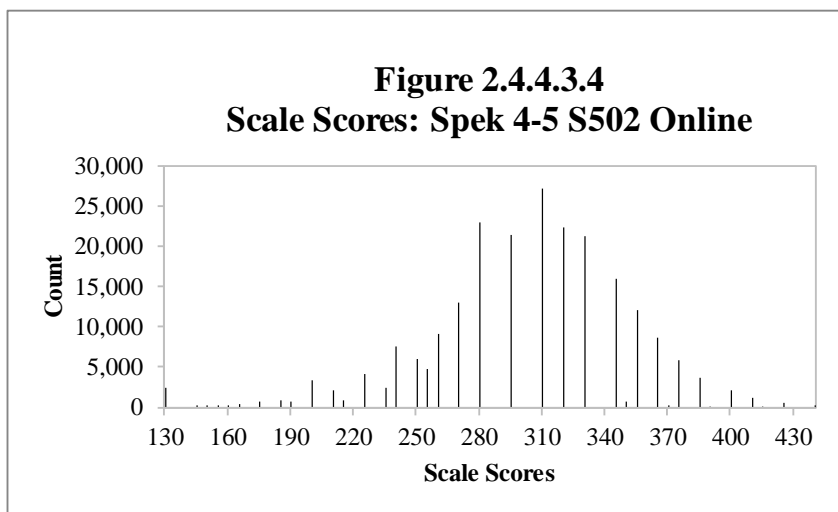


Table 2.4.4.3.4

Scale Score Descriptive Statistics: Spek 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	125,763	130	443	303.64	48.56
5	99,749	130	443	302.98	49.74
Total	225,512	130	443	303.35	49.09



2.4.4.4 Grades 6–8

Table 2.4.4.4.1

Scale Score Descriptive Statistics: Spek 6-8 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	1,275	148	214	200.22	22.57
7	1,661	148	214	198.66	23.34
8	2,828	214	200.95	21.94	
Total	5,764	148	214	200.13	22.51

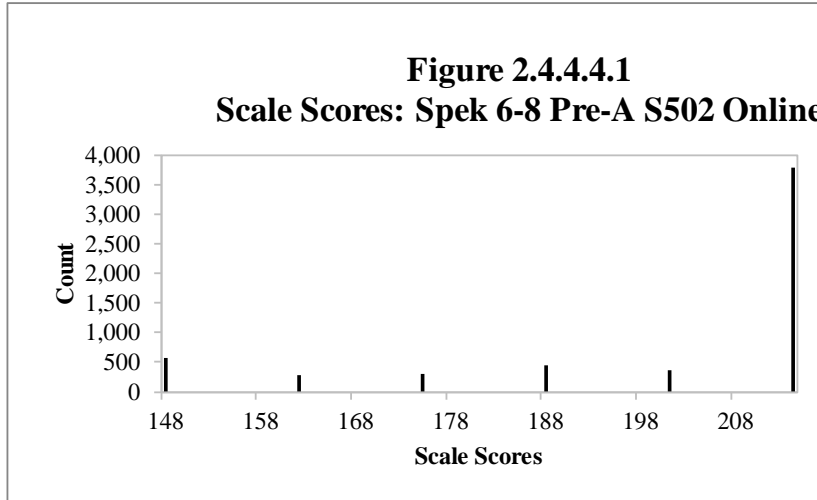


Table 2.4.4.4.2

Scale Score Descriptive Statistics: Spek 6-8 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	13,115	148	416	270.87	48.64
7	11,744	148	416	267.25	50.28
8	21,847	148	437	289.34	50.80
Total	46,706	148	437	278.60	51.09

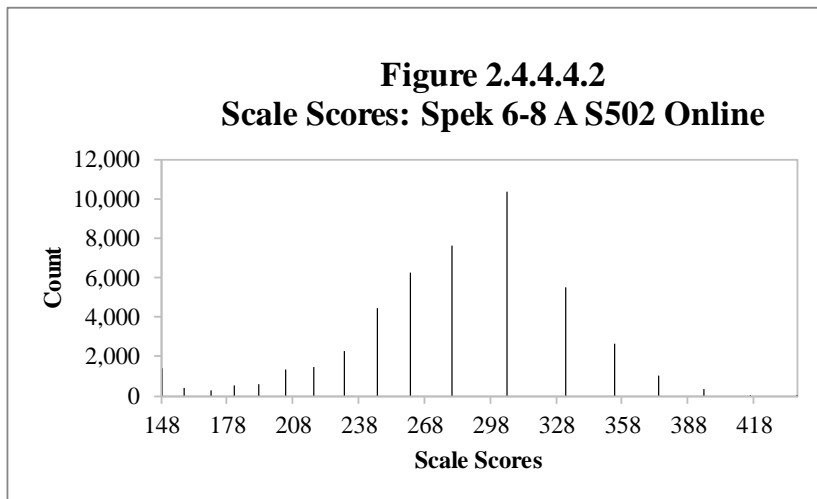


Table 2.4.4.4.3

Scale Score Descriptive Statistics: Spek 6-8 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	66,866	148	463	318.17	42.32
7	67,981	148	463	323.93	43.59
8	48,449	148	463	334.34	42.82
Total	183,296	148	463	324.58	43.39

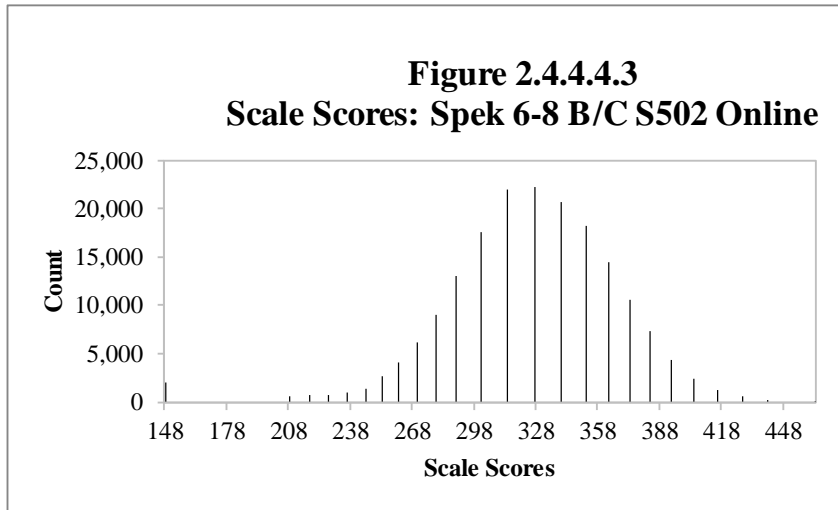
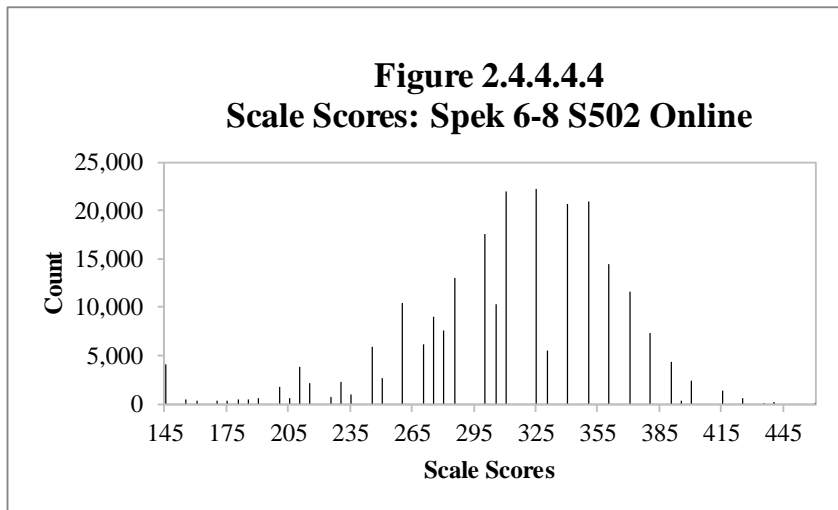


Table 2.4.4.4.4

Scale Score Descriptive Statistics: Spek 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	81,256	148	463	308.69	48.50
7	81,386	148	463	313.19	51.30
8	73,124	148	463	315.74	54.33
Total	235,766	148	463	312.43	51.41



2.4.4.5 Grades 9–12

Table 2.4.4.5.1

Scale Score Descriptive Statistics: Spek 9-12 Pre-A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	2,366	172	224	209.94	19.76
10	3,951	172	224	214.57	17.07
11	3,125	172	224	217.44	14.78
12	2,546	172	224	218.35	14.39
Total	11,988	172	224	215.21	16.82

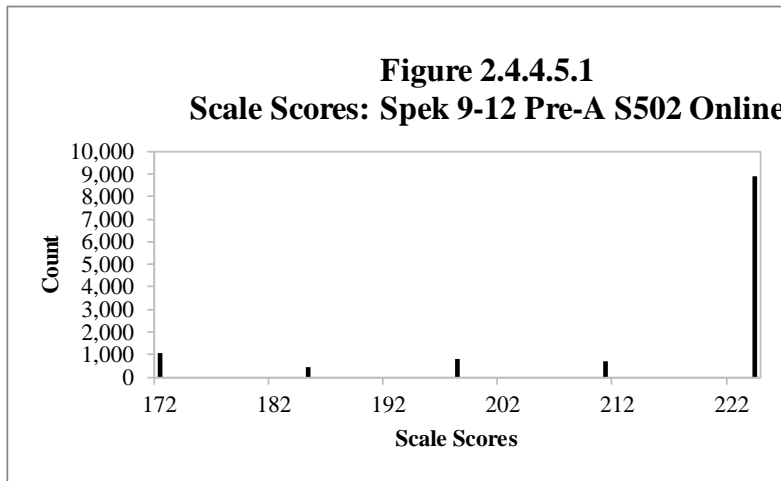


Table 2.4.4.5.2

Scale Score Descriptive Statistics: Spek 9-12 A S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	30,113	172	449	291.21	48.58
10	20,130	172	428	288.41	46.63
11	7,987	172	449	285.19	45.24
12	14,205	172	449	306.71	48.26
Total	72,435	172	449	292.81	48.15

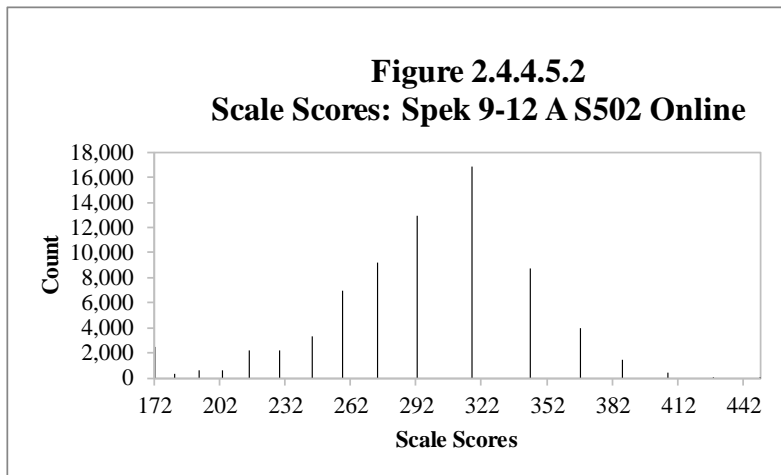


Table 2.4.4.5.3

Scale Score Descriptive Statistics: Spek 9-12 B/C S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	29,229	172	476	338.13	39.40
10	30,737	172	476	338.17	41.07
11	33,225	172	476	335.62	43.79
12	17,495	172	476	345.05	42.03
Total	110,686	172	476	338.48	41.74

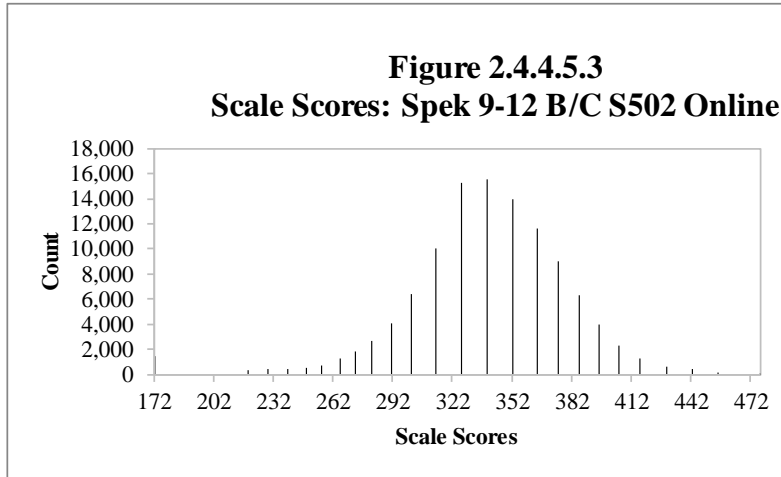
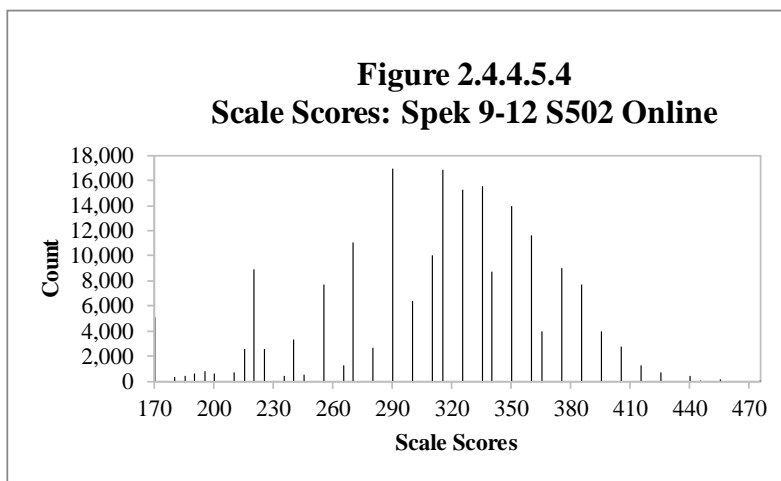


Table 2.4.4.5.4

Scale Score Descriptive Statistics: Spek 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	61,708	172	476	310.32	53.22
10	54,818	172	476	310.99	55.11
11	44,337	172	476	318.20	54.41
12	34,246	172	476	319.73	55.19
Total	195,109	172	476	313.95	54.53



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.

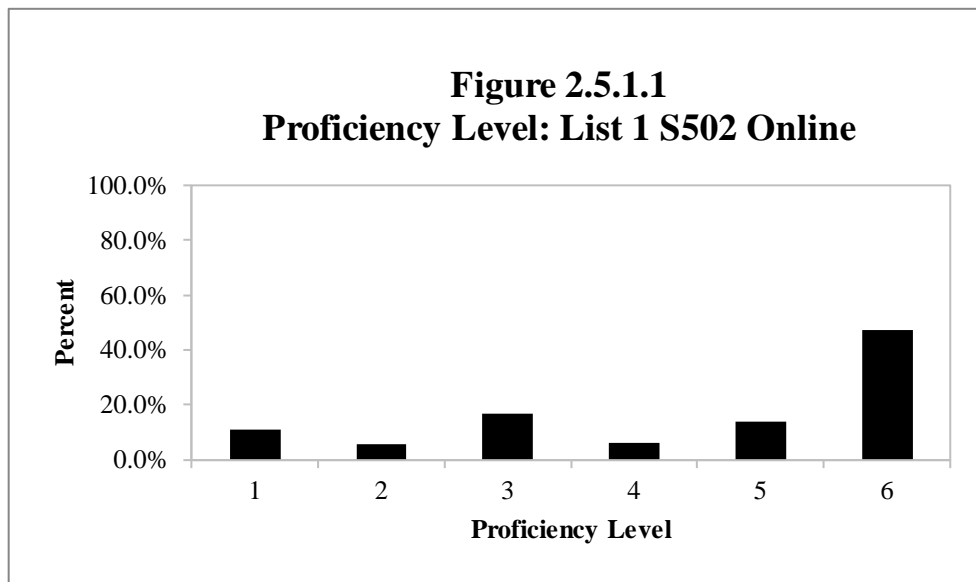
2.5.1 Listening

2.5.1.1 Grade 1

Table 2.5.1.1

Proficiency Level Distribution: List 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	13,866	10.82%	13,866	10.82%
2	7,497	5.85%	7,497	5.85%
3	21,190	16.53%	21,190	16.53%
4	7,598	5.93%	7,598	5.93%
5	17,561	13.70%	17,561	13.70%
6	60,476	47.18%	60,476	47.18%
Total	128,188	100.00%	128,188	100.00%

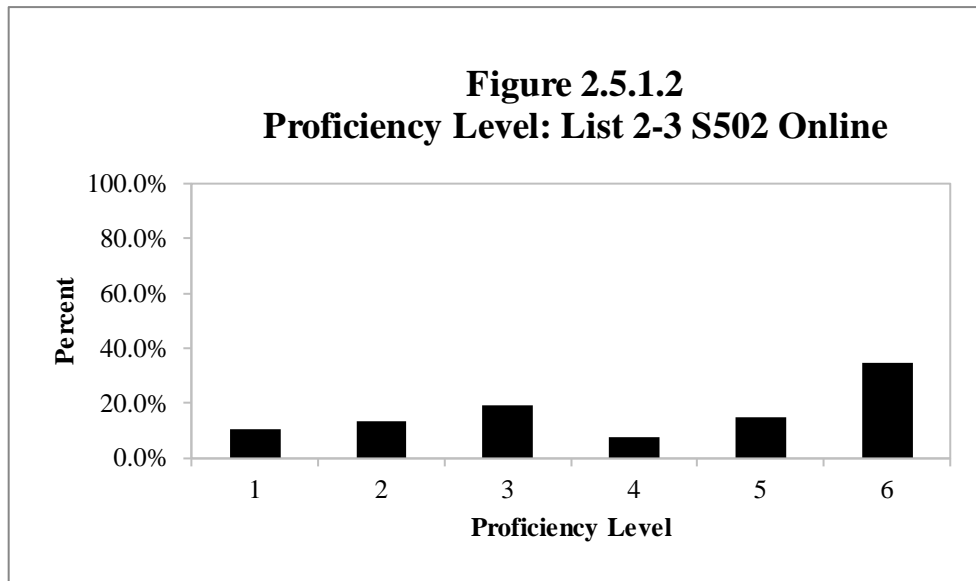


2.5.1.2 Grades 2–3

Table 2.5.1.2

Proficiency Level Distribution: List 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	13,328	10.89%	12,382	10.13%	25,710	10.51%
2	17,275	14.12%	15,117	12.36%	32,392	13.24%
3	26,363	21.54%	20,600	16.85%	46,963	19.20%
4	9,632	7.87%	9,087	7.43%	18,719	7.65%
5	14,881	12.16%	21,582	17.65%	36,463	14.90%
6	40,889	33.41%	43,516	35.59%	84,405	34.50%
Total	122,368	100.00%	122,284	100.00%	244,652	100.00%

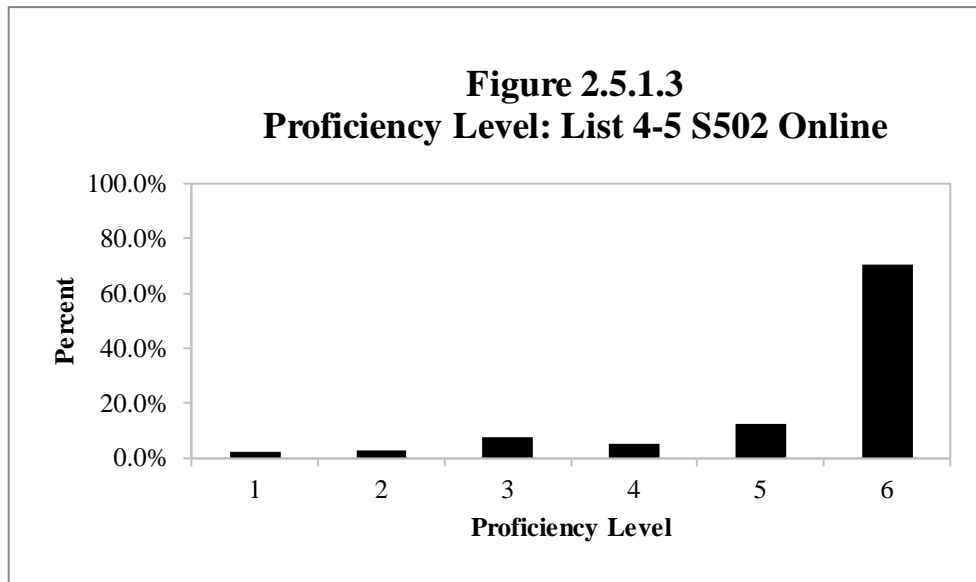


2.5.1.3 Grades 4–5

Table 2.5.1.3

Proficiency Level Distribution: List 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	2,146	1.70%	2,524	2.53%	4,670	2.07%
2	3,097	2.46%	3,049	3.05%	6,146	2.72%
3	7,892	6.26%	8,724	8.74%	16,616	7.36%
4	5,349	4.24%	5,781	5.79%	11,130	4.93%
5	12,920	10.25%	15,287	15.32%	28,207	12.49%
6	94,609	75.08%	64,446	64.57%	159,055	70.43%
Total	126,013	100.00%	99,811	100.00%	225,824	100.00%

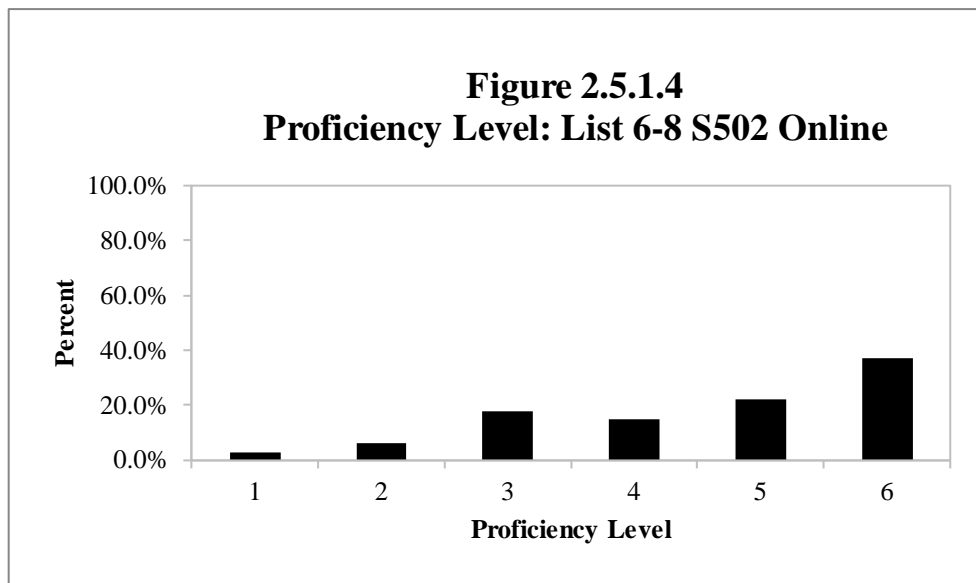


2.5.1.4 Grades 6–8

Table 2.5.1.4

Proficiency Level Distribution: List 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	1,407	1.72%	1,739	2.13%	2,887	3.94%	6,033	2.55%
2	3,878	4.75%	4,946	6.06%	5,140	7.01%	13,964	5.90%
3	12,890	15.80%	15,596	19.12%	13,036	17.78%	41,522	17.56%
4	10,591	12.98%	12,812	15.71%	12,007	16.38%	35,410	14.97%
5	22,508	27.59%	16,008	19.62%	13,776	18.79%	52,292	22.11%
6	30,317	37.16%	30,471	37.35%	26,479	36.11%	87,267	36.90%
Total	81,591	100.00%	81,572	100.00%	73,325	100.00%	236,488	100.00%

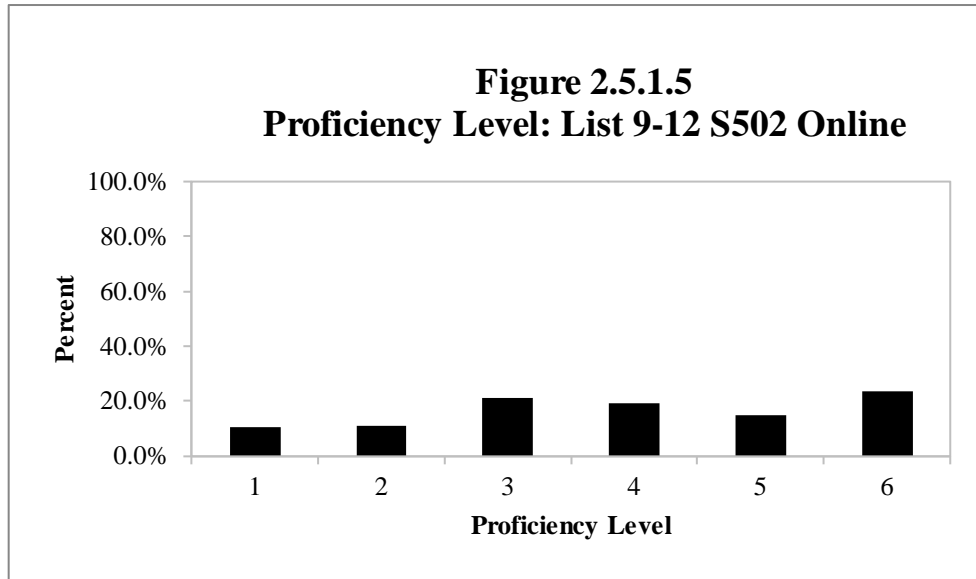


2.5.1.5 Grades 9–12

Table 2.5.1.5

Proficiency Level Distribution: List 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	3,823	6.20%	6,396	11.62%	5,253	11.71%	4,536	13.21%	20,008	10.21%
2	7,978	12.95%	5,472	9.94%	4,194	9.35%	3,939	11.47%	21,583	11.02%
3	13,049	21.18%	12,229	22.22%	8,632	19.24%	6,970	20.30%	40,880	20.87%
4	11,807	19.16%	10,273	18.66%	9,126	20.34%	6,772	19.72%	37,978	19.39%
5	8,447	13.71%	7,907	14.36%	8,132	18.12%	4,600	13.40%	29,086	14.85%
6	16,520	26.81%	12,767	23.19%	9,535	21.25%	7,524	21.91%	46,346	23.66%
Total	61,624	100.00%	55,044	100.00%	44,872	100.00%	34,341	100.00%	195,881	100.00%



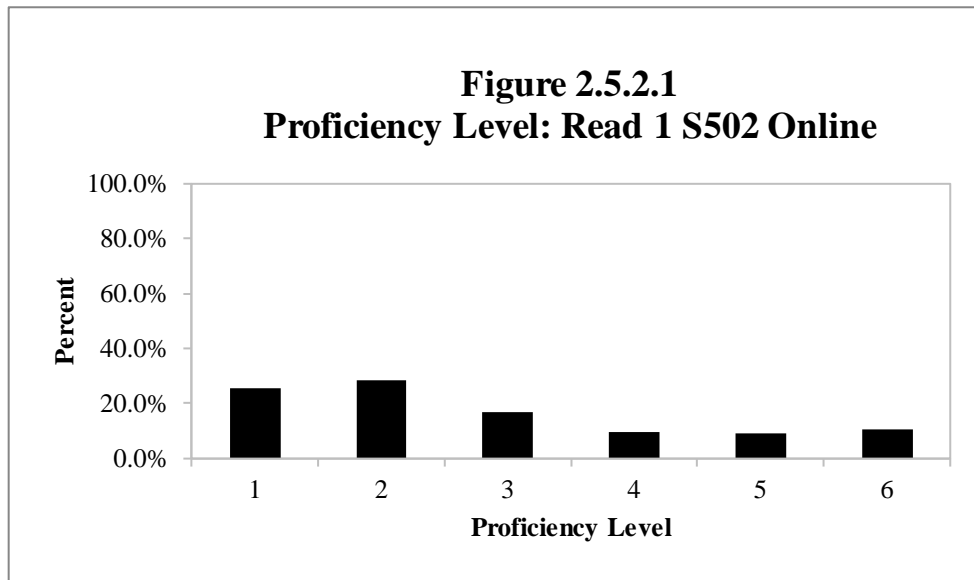
2.5.2 Reading

2.5.2.1 Grade 1

Table 2.5.2.1

Proficiency Level Distribution: Read 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	33,633	25.66%	33,633	25.66%
2	37,404	28.54%	37,404	28.54%
3	22,139	16.89%	22,139	16.89%
4	12,675	9.67%	12,675	9.67%
5	11,764	8.97%	11,764	8.97%
6	13,463	10.27%	13,463	10.27%
Total	131,078	100.00%	131,078	100.00%

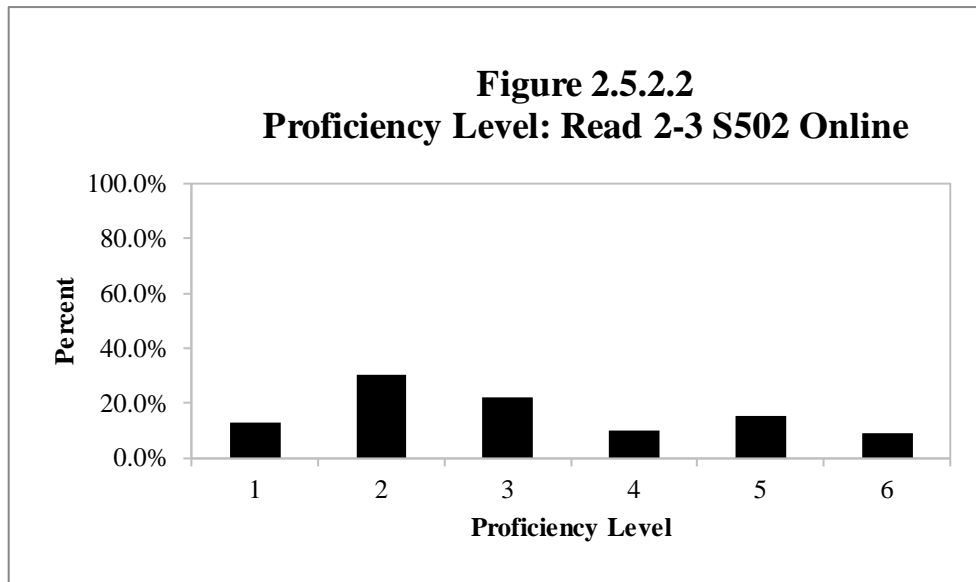


2.5.2.2 Grades 2–3

Table 2.5.2.2

Proficiency Level Distribution: Read 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	12,233	9.88%	19,501	15.84%	31,734	12.85%
2	32,493	26.24%	42,435	34.47%	74,928	30.34%
3	33,896	27.37%	20,935	17.01%	54,831	22.20%
4	15,045	12.15%	9,748	7.92%	24,793	10.04%
5	20,721	16.73%	17,390	14.13%	38,111	15.43%
6	9,447	7.63%	13,090	10.63%	22,537	9.13%
Total	123,835	100.00%	123,099	100.00%	246,934	100.00%

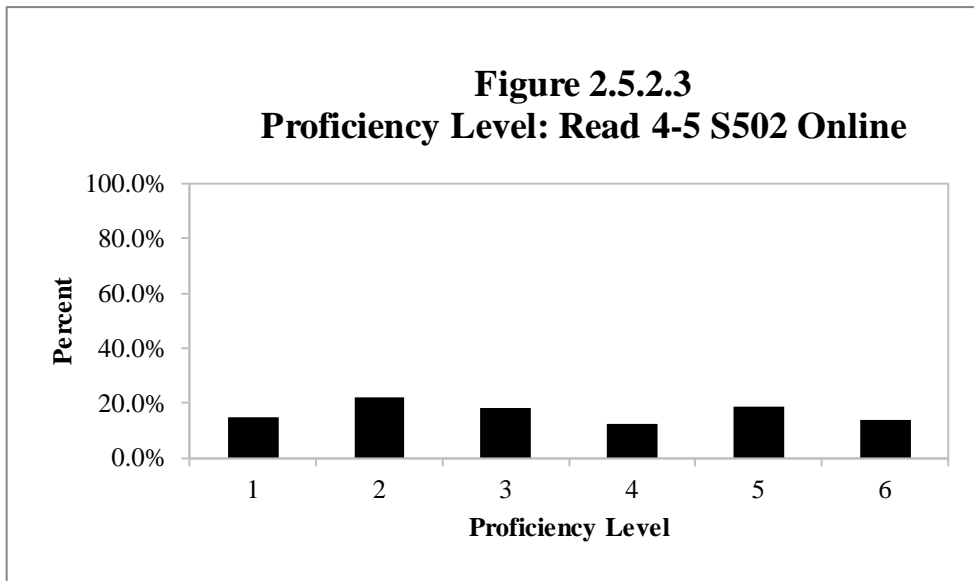


2.5.2.3 Grades 4–5

Table 2.5.2.3

Proficiency Level Distribution: Read 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	16,983	13.52%	16,661	16.74%	33,644	14.94%
2	25,862	20.58%	23,852	23.97%	49,714	22.08%
3	20,987	16.70%	19,463	19.56%	40,450	17.96%
4	18,731	14.91%	9,630	9.68%	28,361	12.60%
5	24,870	19.79%	17,157	17.24%	42,027	18.66%
6	18,219	14.50%	12,751	12.81%	30,970	13.75%
Total	125,652	100.00%	99,514	100.00%	225,166	100.00%

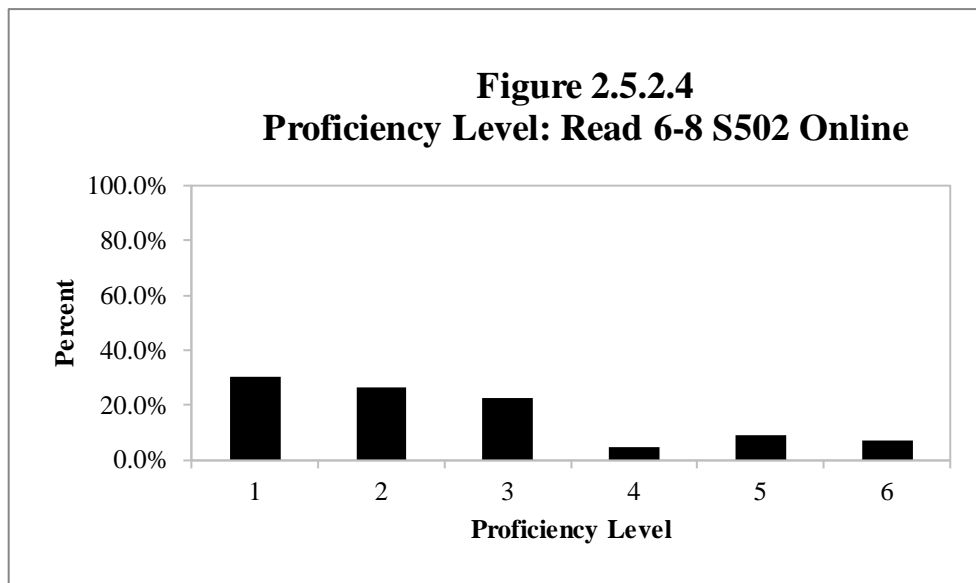


2.5.2.4 Grades 6–8

Table 2.5.2.4

Proficiency Level Distribution: Read 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	25,161	30.76%	23,753	29.17%	22,998	31.48%	71,912	30.44%
2	21,298	26.04%	21,980	26.99%	18,848	25.80%	62,126	26.30%
3	21,591	26.40%	18,038	22.15%	13,864	18.98%	53,493	22.64%
4	4,265	5.21%	3,453	4.24%	2,970	4.07%	10,688	4.52%
5	6,376	7.80%	7,848	9.64%	6,901	9.45%	21,125	8.94%
6	3,095	3.78%	6,352	7.80%	7,468	10.22%	16,915	7.16%
Total	81,786	100.00%	81,424	100.00%	73,049	100.00%	236,259	100.00%

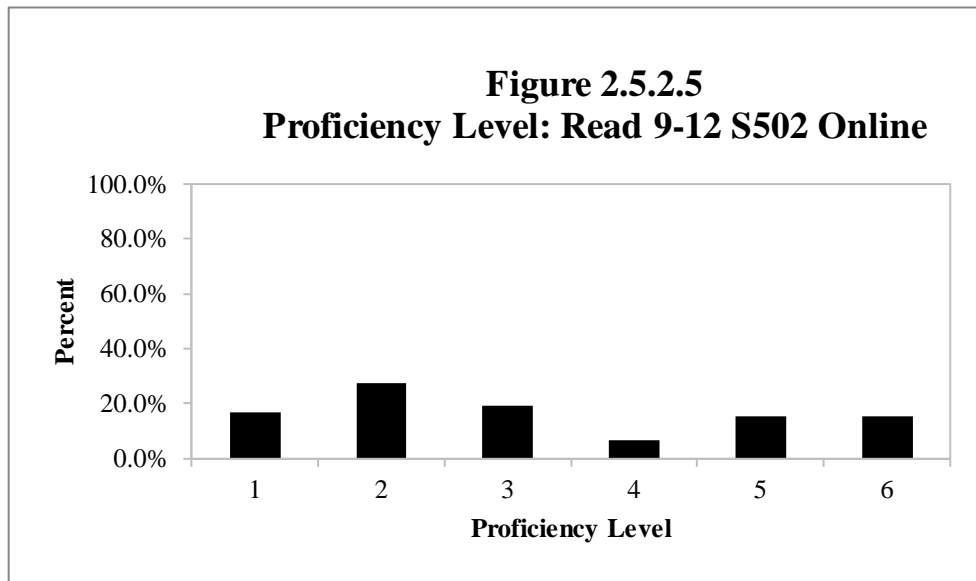


2.5.2.5 Grades 9–12

Table 2.5.2.5

Proficiency Level Distribution: Read 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	9,257	15.17%	9,764	17.93%	7,364	16.64%	5,697	16.75%	32,082	16.56%
2	16,413	26.89%	14,358	26.37%	12,371	27.95%	10,097	29.69%	53,239	27.48%
3	13,143	21.53%	10,480	19.25%	7,743	17.49%	5,722	16.82%	37,088	19.14%
4	4,031	6.60%	3,643	6.69%	2,765	6.25%	2,450	7.20%	12,889	6.65%
5	9,020	14.78%	7,828	14.38%	6,975	15.76%	5,432	15.97%	29,255	15.10%
6	9,174	15.03%	8,382	15.39%	7,048	15.92%	4,611	13.56%	29,215	15.08%
Total	61,038	100.00%	54,455	100.00%	44,266	100.00%	34,009	100.00%	193,768	100.00%



2.5.3 Writing

2.5.3.1 Grade 1

Table 2.5.3.1.1

Proficiency Level Distribution: Writ 1 A S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	59,616	51.55%	59,616	51.55%
2	49,864	43.11%	49,864	43.11%
3	6,157	5.32%	6,157	5.32%
4	18	0.02%	18	0.02%
5	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%
Total	115,655	100.00%	115,655	100.00%

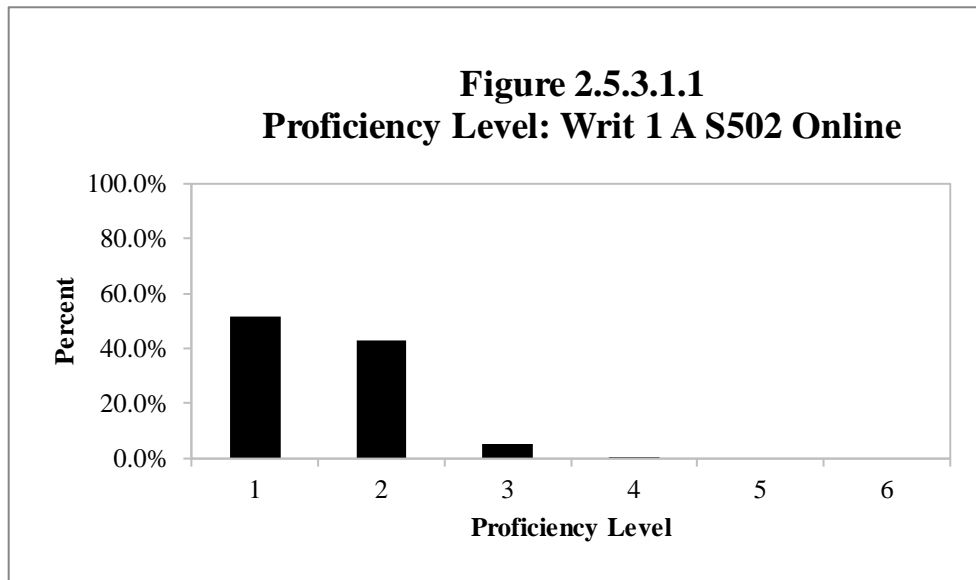


Table 2.5.3.1.2

Proficiency Level Distribution: Writ 1 B/C S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	927	4.31%	927	4.31%
2	6,521	30.30%	6,521	30.30%
3	13,112	60.94%	13,112	60.94%
4	948	4.41%	948	4.41%
5	7	0.03%	7	0.03%
6	3	0.01%	3	0.01%
Total	21,518	100.00%	21,518	100.00%

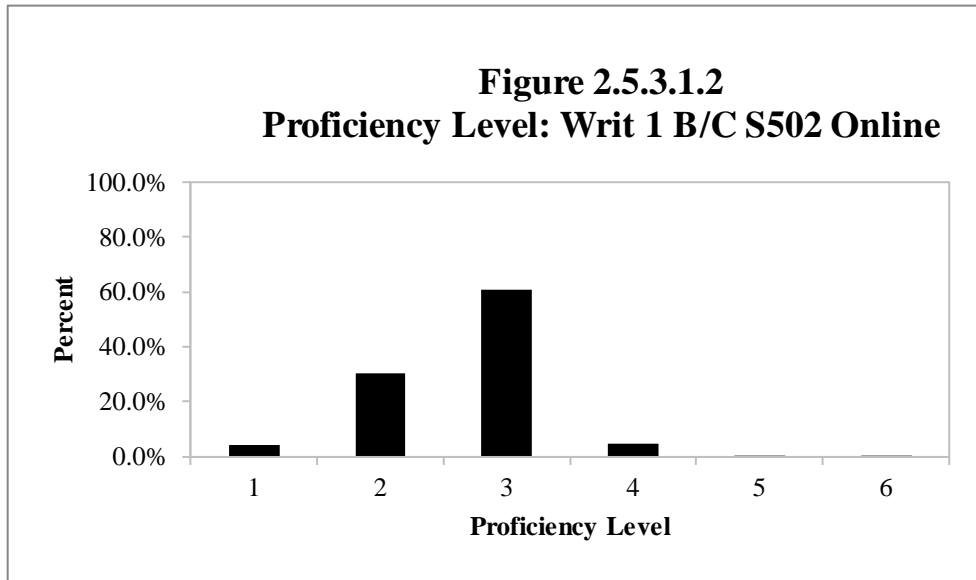
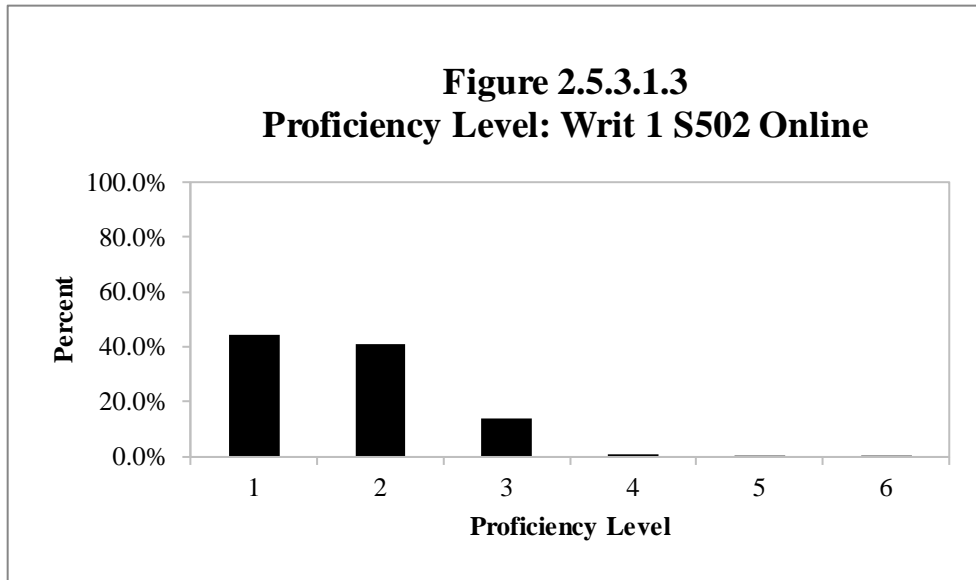


Table 2.5.3.1.3

Proficiency Level Distribution: Writ 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	60,543	44.14%	60,543	44.14%
2	56,385	41.11%	56,385	41.11%
3	19,269	14.05%	19,269	14.05%
4	966	0.70%	966	0.70%
5	7	0.01%	7	0.01%
6	3	0.00%	3	0.00%
Total	137,173	100.00%	137,173	100.00%



2.5.3.2 Grades 2–3

Table 2.5.3.2.1

Proficiency Level Distribution: Writ 2-3 A S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	15,275	36.44%	8,416	26.32%	23,691	32.06%
2	16,049	38.28%	10,618	33.21%	26,667	36.09%
3	10,371	24.74%	12,780	39.97%	23,151	31.33%
4	225	0.54%	158	0.49%	383	0.52%
5	0	0.00%	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%	0	0.00%
Total	41,920	100.00%	31,972	100.00%	73,892	100.00%

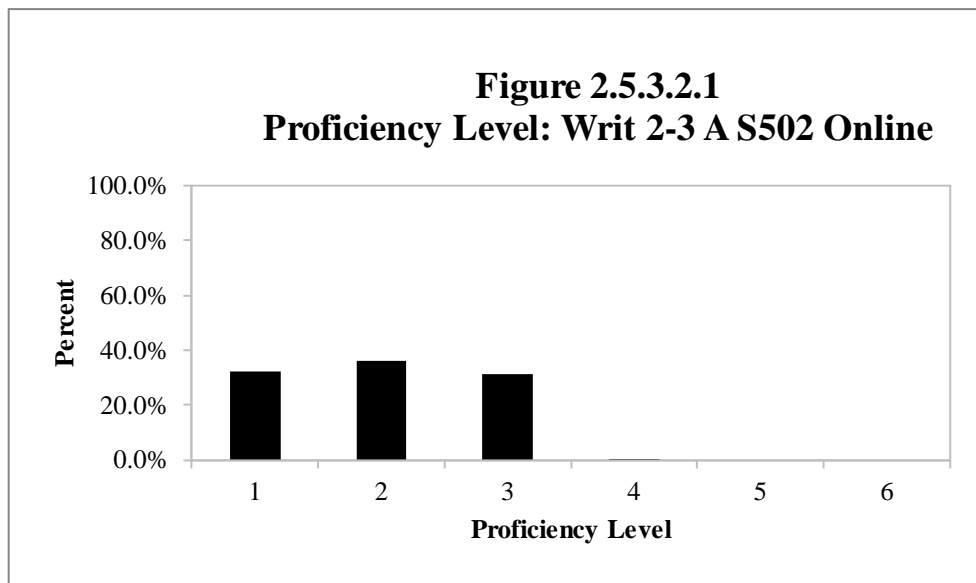


Table 2.5.3.2.2

Proficiency Level Distribution: Writ 2-3 B/C S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	5,246	5.89%	1,797	1.83%	7,043	3.76%
2	18,686	20.97%	6,981	7.11%	25,667	13.70%
3	53,612	60.17%	56,954	57.97%	110,566	59.02%
4	11,548	12.96%	32,443	33.02%	43,991	23.48%
5	5	0.01%	54	0.05%	59	0.03%
6	1	0.00%	10	0.01%	11	0.01%
Total	89,098	100.00%	98,239	100.00%	187,337	100.00%

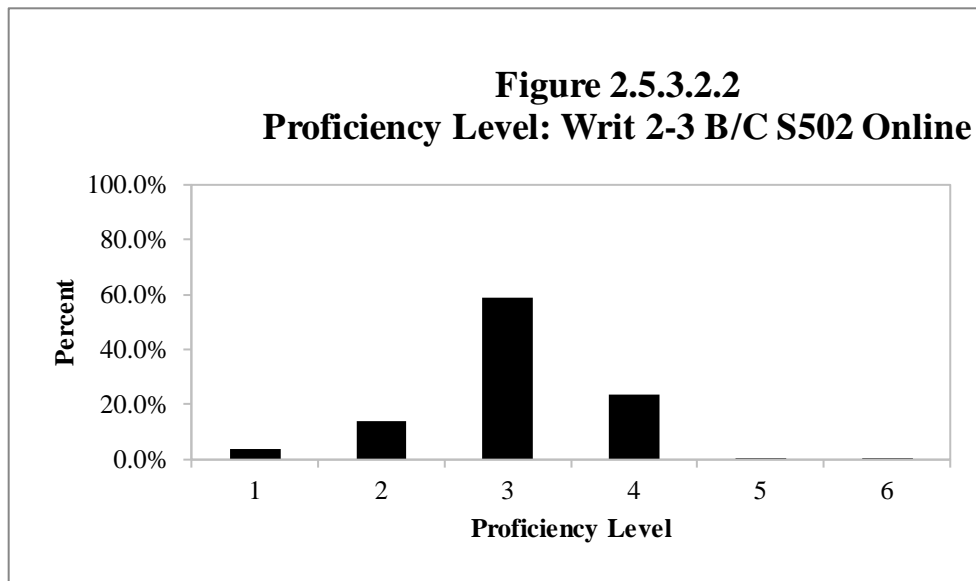
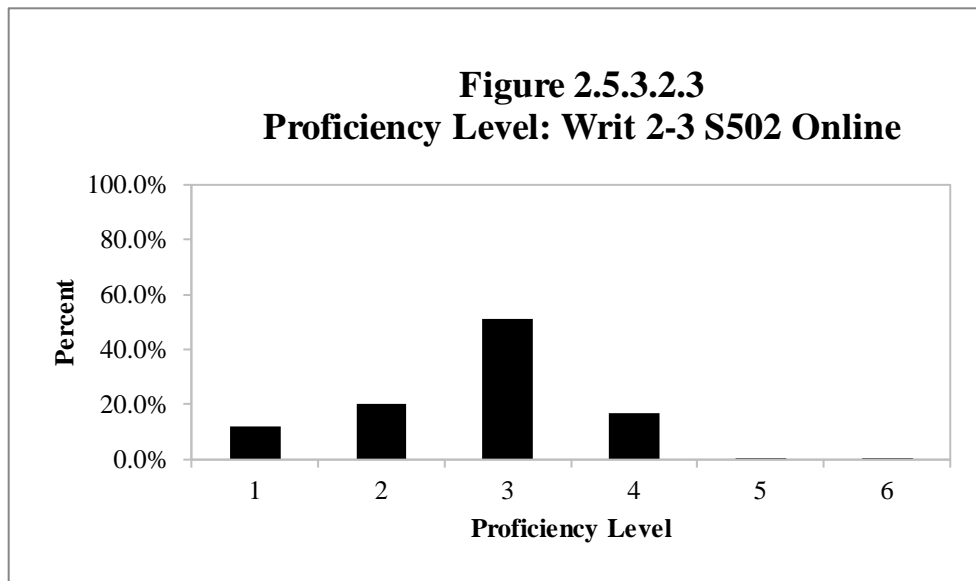


Table 2.5.3.2.3

Proficiency Level Distribution: Writ 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	20,521	15.66%	10,213	7.84%	30,734	11.77%
2	34,735	26.51%	17,599	13.52%	52,334	20.03%
3	63,983	48.84%	69,734	53.55%	133,717	51.19%
4	11,773	8.99%	32,601	25.04%	44,374	16.99%
5	5	0.00%	54	0.04%	59	0.02%
6	1	0.00%	10	0.01%	11	0.00%
Total	131,018	100.00%	130,211	100.00%	261,229	100.00%



2.5.3.3 Grades 4–5

Table 2.5.3.3.1

Proficiency Level Distribution: Writ 4-5 A S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	8,321	35.34%	6,014	26.76%	14,335	31.15%
2	4,780	20.30%	7,181	31.95%	11,961	25.99%
3	10,031	42.60%	8,335	37.09%	18,366	39.91%
4	413	1.75%	942	4.19%	1,355	2.94%
5	2	0.01%	3	0.01%	5	0.01%
6	0	0.00%	0	0.00%	0	0.00%
Total	23,547	100.00%	22,475	100.00%	46,022	100.00%

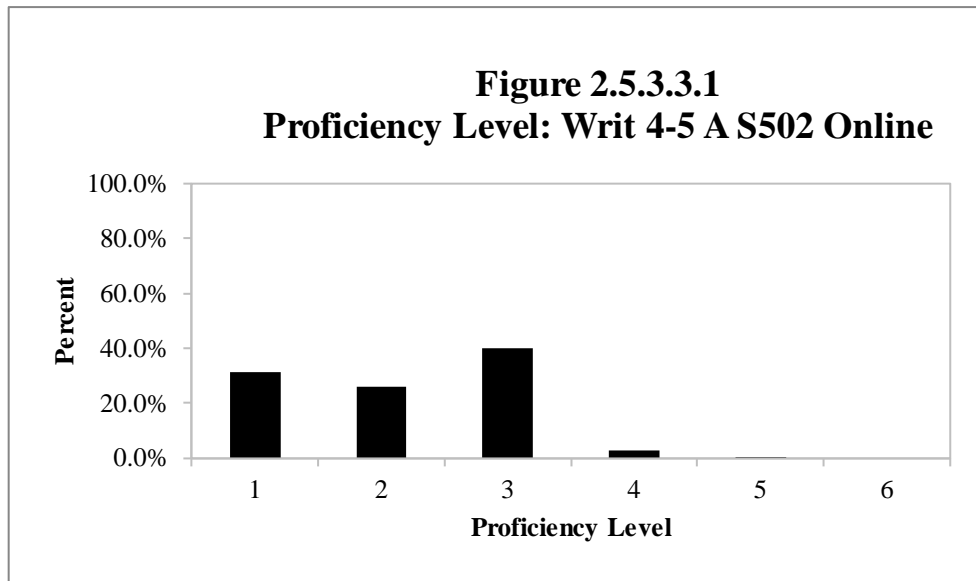


Table 2.5.3.3.2

Proficiency Level Distribution: Writ 4-5 B/C S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	1,045	0.98%	266	0.33%	1,311	0.70%
2	669	0.62%	512	0.63%	1,181	0.63%
3	69,050	64.43%	40,249	49.77%	109,299	58.12%
4	34,109	31.82%	35,942	44.45%	70,051	37.25%
5	1,693	1.58%	3,575	4.42%	5,268	2.80%
6	612	0.57%	321	0.40%	933	0.50%
Total	107,178	100.00%	80,865	100.00%	188,043	100.00%

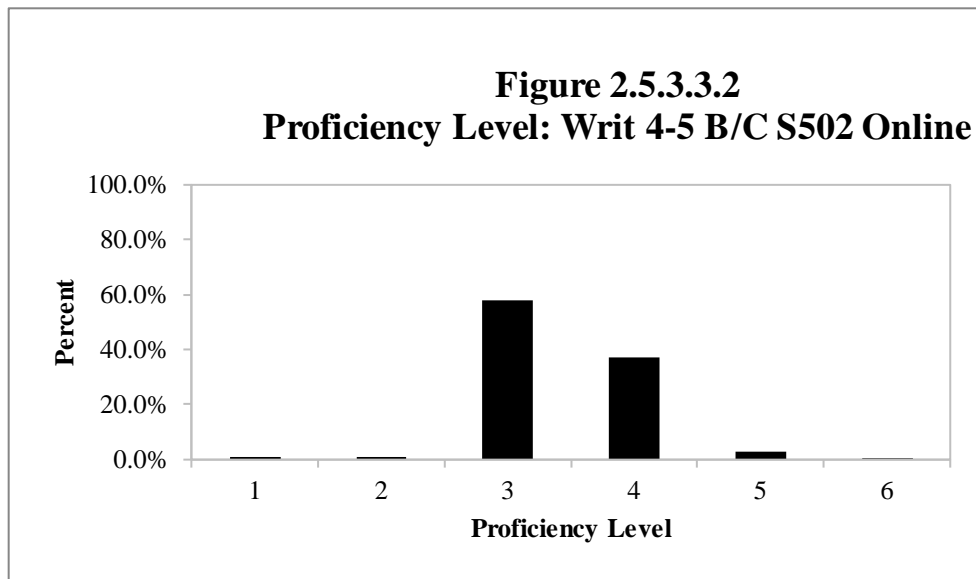
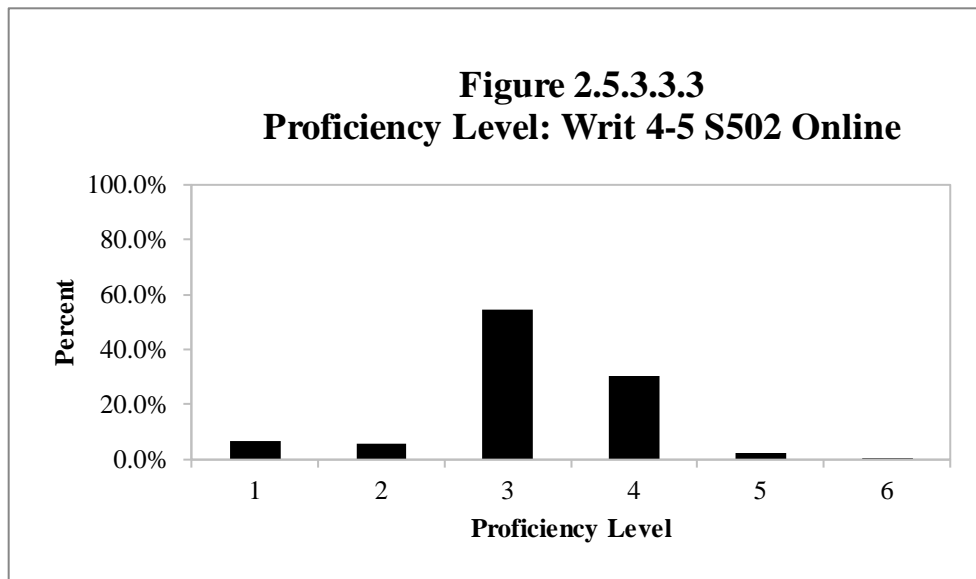


Table 2.5.3.3.3

Proficiency Level Distribution: Writ 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	9,366	7.16%	6,280	6.08%	15,646	6.68%
2	5,449	4.17%	7,693	7.44%	13,142	5.61%
3	79,081	60.49%	48,584	47.01%	127,665	54.54%
4	34,522	26.41%	36,884	35.69%	71,406	30.51%
5	1,695	1.30%	3,578	3.46%	5,273	2.25%
6	612	0.47%	321	0.31%	933	0.40%
Total	130,725	100.00%	103,340	100.00%	234,065	100.00%



2.5.3.4 Grades 6–8

Table 2.5.3.4.1

Proficiency Level Distribution: Writ 6-8 A S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	5,452	20.95%	6,492	20.90%	7,996	25.38%	19,940	22.51%
2	8,082	31.06%	13,188	42.45%	9,360	29.71%	30,630	34.58%
3	12,267	47.14%	10,740	34.57%	13,849	43.96%	36,856	41.60%
4	219	0.84%	644	2.07%	296	0.94%	1,159	1.31%
5	0	0.00%	0	0.00%	2	0.01%	2	0.00%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	26,020	100.00%	31,064	100.00%	31,503	100.00%	88,587	100.00%

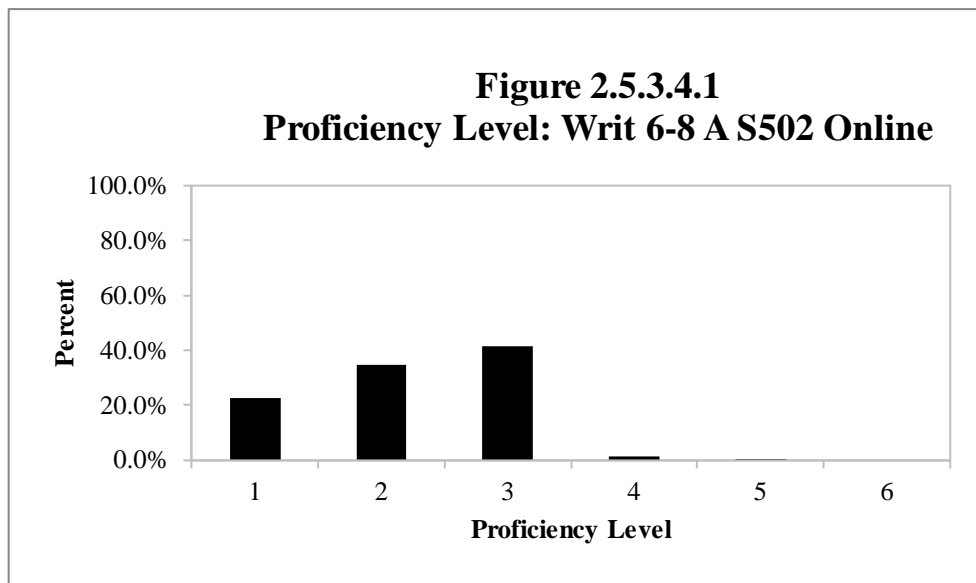


Table 2.5.3.4.2

Proficiency Level Distribution: Writ 6-8 B/C S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	639	1.09%	259	0.49%	283	0.65%	1,181	0.76%
2	7,785	13.23%	7,146	13.39%	3,348	7.65%	18,279	11.72%
3	44,632	75.85%	34,619	64.89%	34,078	77.82%	113,329	72.65%
4	5,716	9.71%	11,255	21.10%	5,968	13.63%	22,939	14.71%
5	69	0.12%	70	0.13%	107	0.24%	246	0.16%
6	3	0.01%	2	0.00%	7	0.02%	12	0.01%
Total	58,844	100.00%	53,351	100.00%	43,791	100.00%	155,986	100.00%

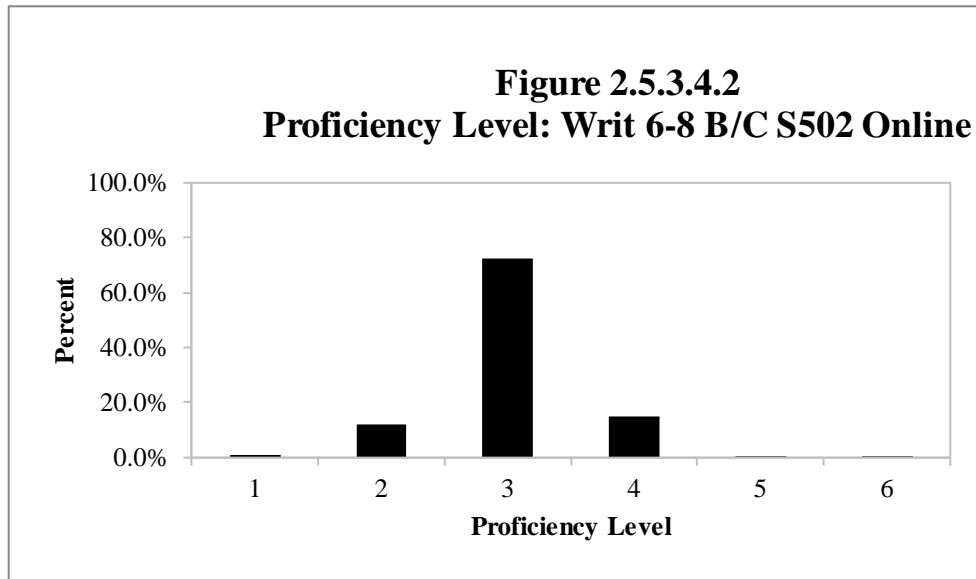
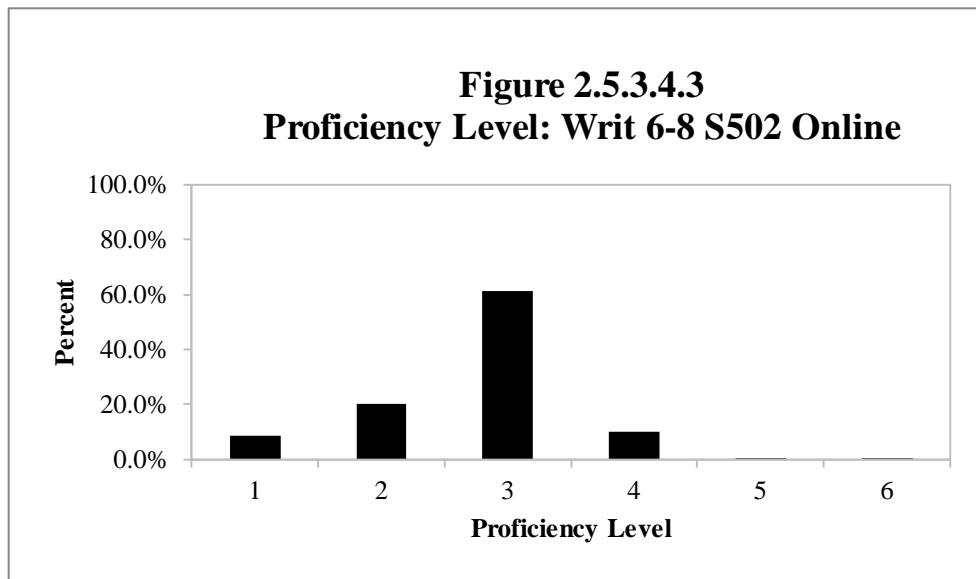


Table 2.5.3.4.3

Proficiency Level Distribution: Writ 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	6,091	7.18%	6,751	8.00%	8,279	11.00%	21,121	8.64%
2	15,867	18.70%	20,334	24.09%	12,708	16.88%	48,909	20.00%
3	56,899	67.05%	45,359	53.73%	47,927	63.65%	150,185	61.41%
4	5,935	6.99%	11,899	14.10%	6,264	8.32%	24,098	9.85%
5	69	0.08%	70	0.08%	109	0.14%	248	0.10%
6	3	0.00%	2	0.00%	7	0.01%	12	0.00%
Total	84,864	100.00%	84,415	100.00%	75,294	100.00%	244,573	100.00%



2.5.3.5 Grades 9–12

Table 2.5.3.5.1

Proficiency Level Distribution: Writ 9-12 A S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	4,163	19.29%	4,192	22.09%	4,158	30.24%	2,280	25.12%	14,793	23.34%
2	7,339	34.01%	6,261	33.00%	4,047	29.43%	2,752	30.32%	20,399	32.19%
3	7,813	36.21%	7,751	40.85%	4,766	34.66%	3,405	37.51%	23,735	37.45%
4	2,259	10.47%	758	4.00%	772	5.61%	639	7.04%	4,428	6.99%
5	5	0.02%	11	0.06%	8	0.06%	1	0.01%	25	0.04%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	21,579	100.00%	18,973	100.00%	13,751	100.00%	9,077	100.00%	63,380	100.00%

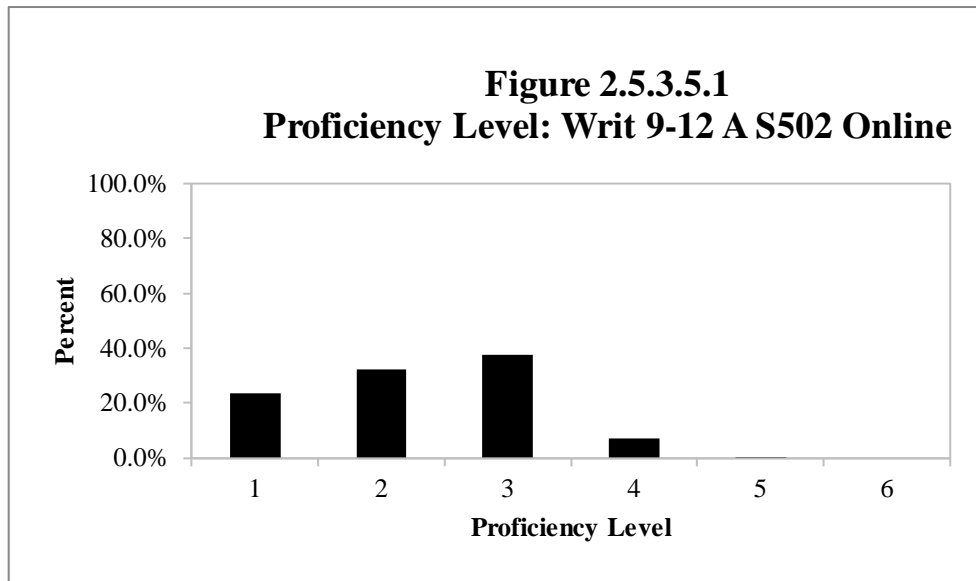


Table 2.5.3.5.2

Proficiency Level Distribution: Writ 9-12 B/C S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	327	0.77%	658	1.74%	1,739	5.37%	1,500	5.71%	4,224	3.05%
2	5,950	14.06%	4,571	12.11%	6,706	20.70%	5,362	20.41%	22,589	16.28%
3	24,448	57.76%	26,719	70.81%	17,951	55.42%	14,549	55.38%	83,667	60.31%
4	11,349	26.81%	5,433	14.40%	5,634	17.39%	4,787	18.22%	27,203	19.61%
5	240	0.57%	350	0.93%	358	1.11%	69	0.26%	1,017	0.73%
6	12	0.03%	3	0.01%	1	0.00%	3	0.01%	19	0.01%
Total	42,326	100.00%	37,734	100.00%	32,389	100.00%	26,270	100.00%	138,719	100.00%

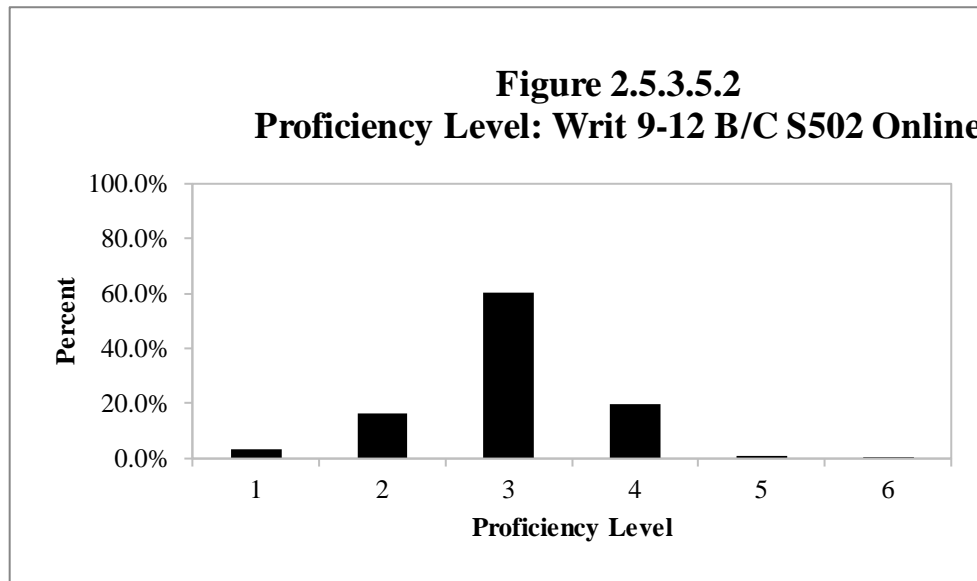
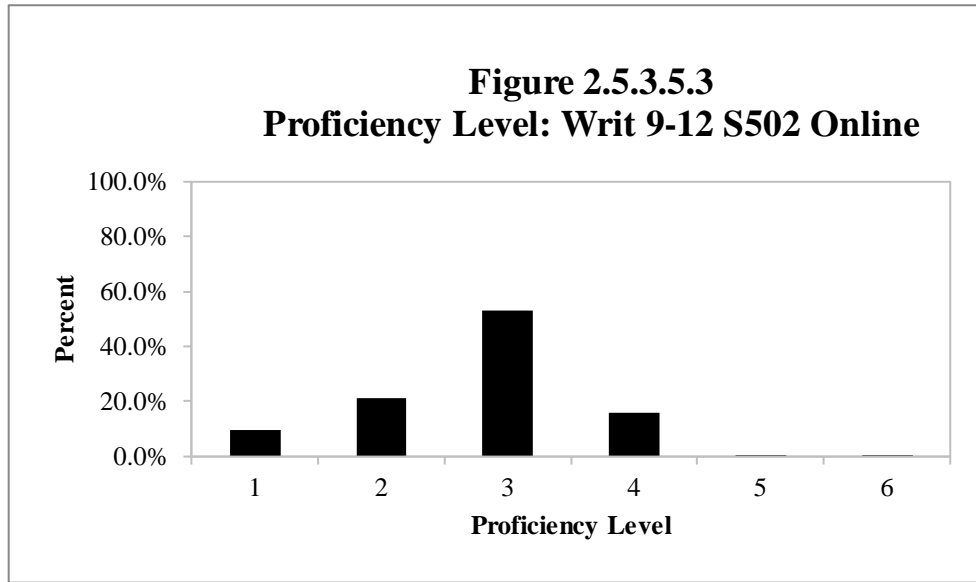


Table 2.5.3.5.3

Proficiency Level Distribution: Writ 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	4,490	7.03%	4,850	8.55%	5,897	12.78%	3,780	10.69%	19,017	9.41%
2	13,289	20.79%	10,832	19.10%	10,753	23.31%	8,114	22.96%	42,988	21.27%
3	32,261	50.48%	34,470	60.79%	22,717	49.23%	17,954	50.79%	107,402	53.14%
4	13,608	21.29%	6,191	10.92%	6,406	13.88%	5,426	15.35%	31,631	15.65%
5	245	0.38%	361	0.64%	366	0.79%	70	0.20%	1,042	0.52%
6	12	0.02%	3	0.01%	1	0.00%	3	0.01%	19	0.01%
Total	63,905	100.00%	56,707	100.00%	46,140	100.00%	35,347	100.00%	202,099	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 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	5,634	100.00%	5,634	100.00%
Total	5,634	100.00%	5,634	100.00%

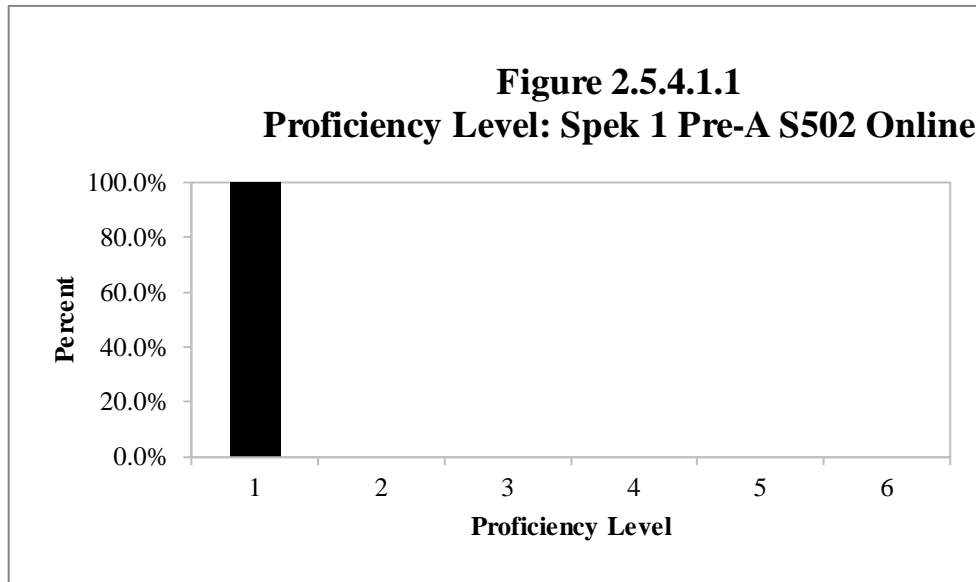


Table 2.5.4.1.2

Proficiency Level Distribution: Spek 1 A S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	12,381	22.27%	12,381	22.27%
2	29,890	53.76%	29,890	53.76%
3	10,289	18.51%	10,289	18.51%
4	3,010	5.41%	3,010	5.41%
5	29	0.05%	29	0.05%
6	0	0.00%	0	0.00%
Total	55,599	100.00%	55,599	100.00%

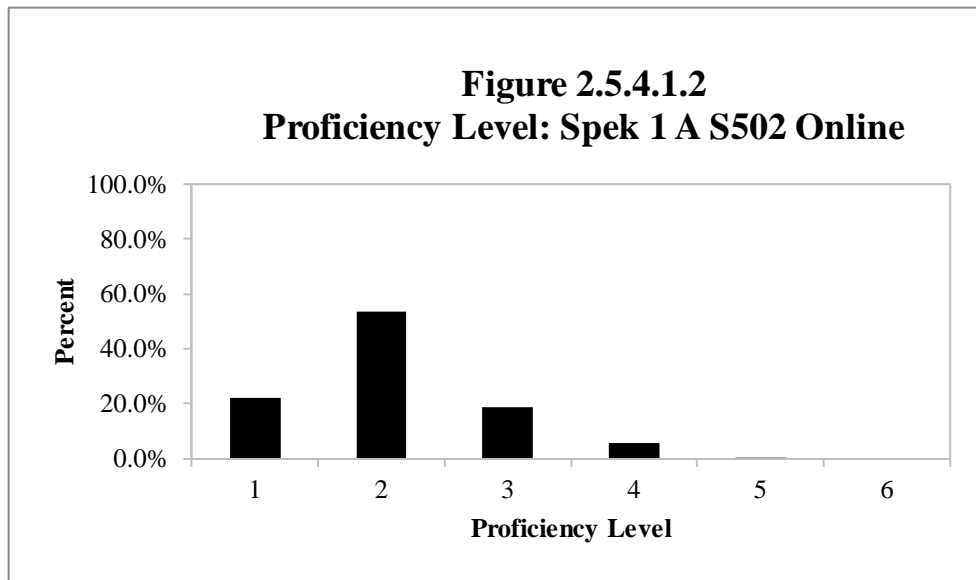


Table 2.5.4.1.3

Proficiency Level Distribution: Spek 1 B/C S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	3,891	5.79%	3,891	5.79%
2	22,285	33.17%	22,285	33.17%
3	27,407	40.80%	27,407	40.80%
4	12,797	19.05%	12,797	19.05%
5	723	1.08%	723	1.08%
6	78	0.12%	78	0.12%
Total	67,181	100.00%	67,181	100.00%

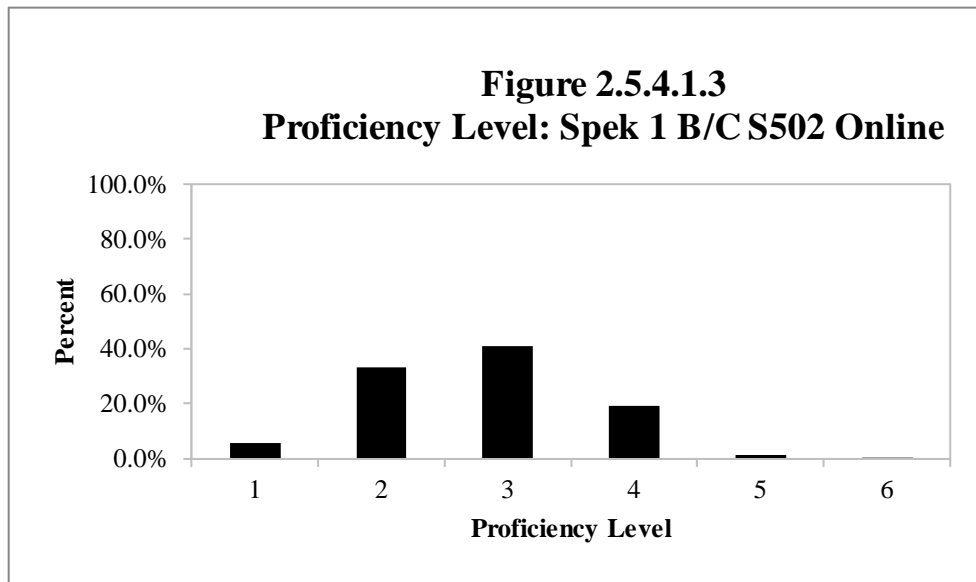
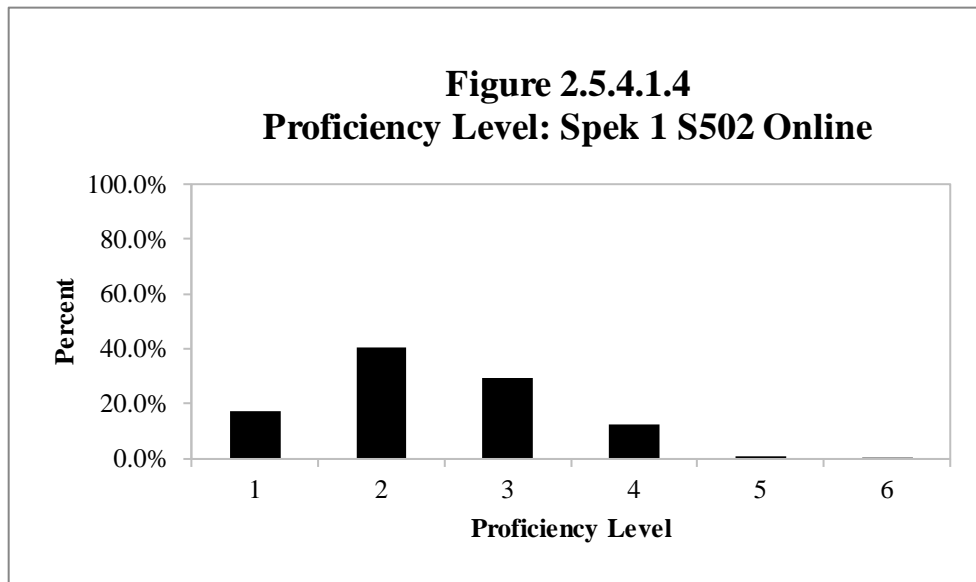


Table 2.5.4.1.4

Proficiency Level Distribution: Spek 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	21,906	17.06%	21,906	17.06%
2	52,175	40.63%	52,175	40.63%
3	37,696	29.36%	37,696	29.36%
4	15,807	12.31%	15,807	12.31%
5	752	0.59%	752	0.59%
6	78	0.06%	78	0.06%
Total	128,414	100.00%	128,414	100.00%



2.5.4.2 Grades 2–3

Table 2.5.4.2.1

Proficiency Level Distribution: Spek 2-3 Pre-A S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	3,050	100.00%	6,101	100.00%	9,151	100.00%
Total	3,050	100.00%	6,101	100.00%	9,151	100.00%

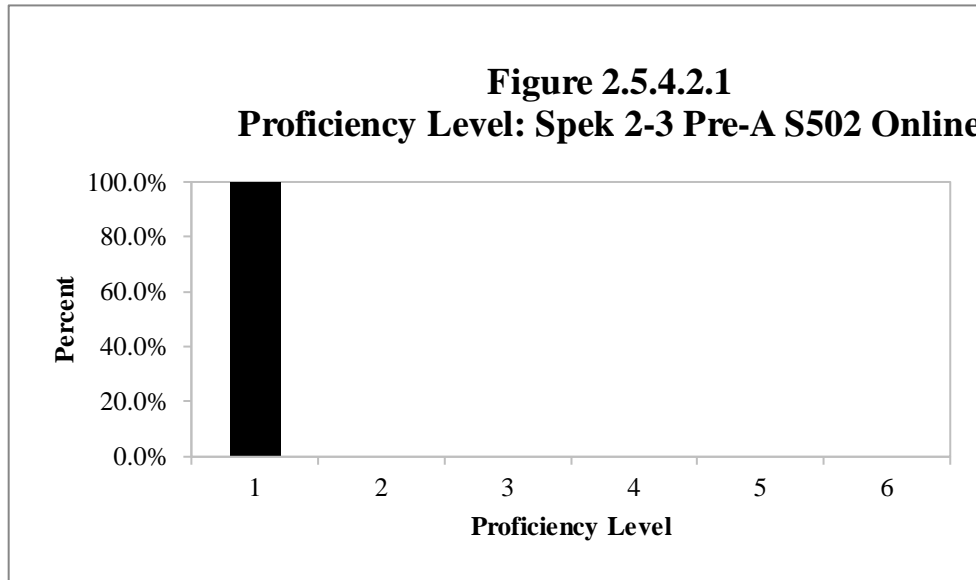


Table 2.5.4.2.2

Proficiency Level Distribution: Spek 2-3 A S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	12,309	35.04%	11,168	33.16%	23,477	34.12%
2	13,085	37.25%	13,866	41.17%	26,951	39.17%
3	9,145	26.03%	7,195	21.36%	16,340	23.75%
4	573	1.63%	1,451	4.31%	2,024	2.94%
5	14	0.04%	0	0.00%	14	0.02%
6	0	0.00%	0	0.00%	0	0.00%
Total	35,126	100.00%	33,680	100.00%	68,806	100.00%

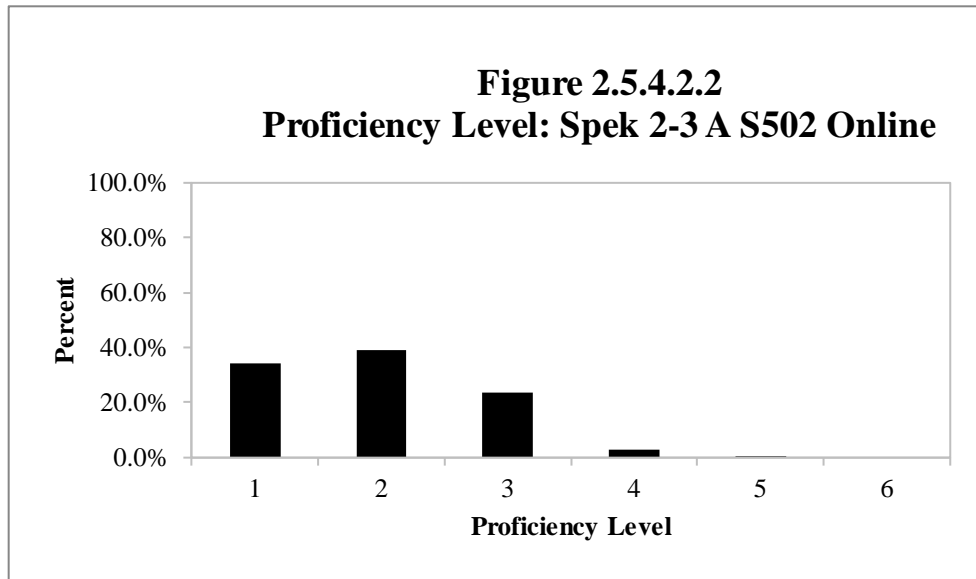


Table 2.5.4.2.3

Proficiency Level Distribution: Spek 2-3 B/C S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	7,132	8.55%	6,676	8.11%	13,808	8.33%
2	32,485	38.95%	16,597	20.17%	49,082	29.62%
3	32,225	38.64%	44,676	54.28%	76,901	46.41%
4	10,557	12.66%	13,287	16.14%	23,844	14.39%
5	891	1.07%	739	0.90%	1,630	0.98%
6	116	0.14%	328	0.40%	444	0.27%
Total	83,406	100.00%	82,303	100.00%	165,709	100.00%

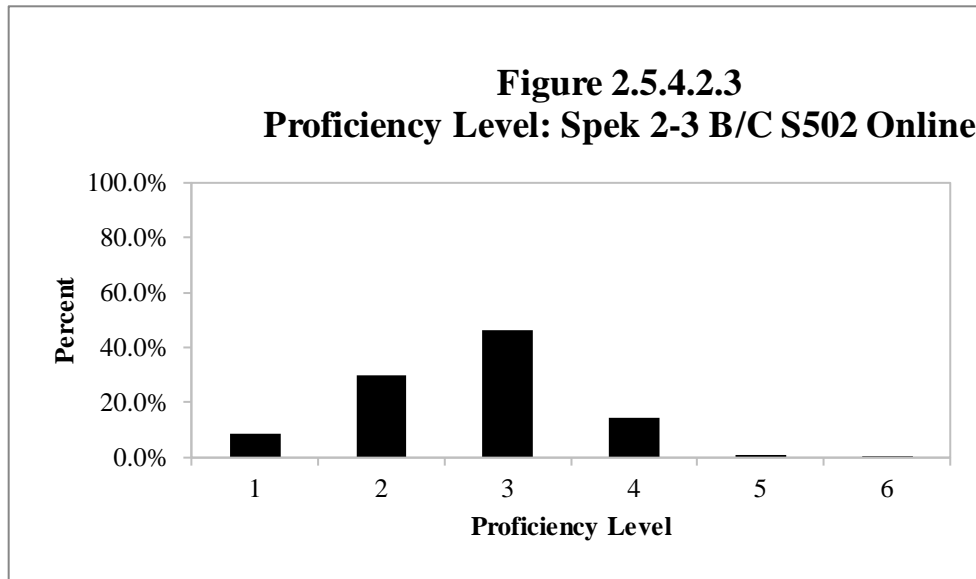
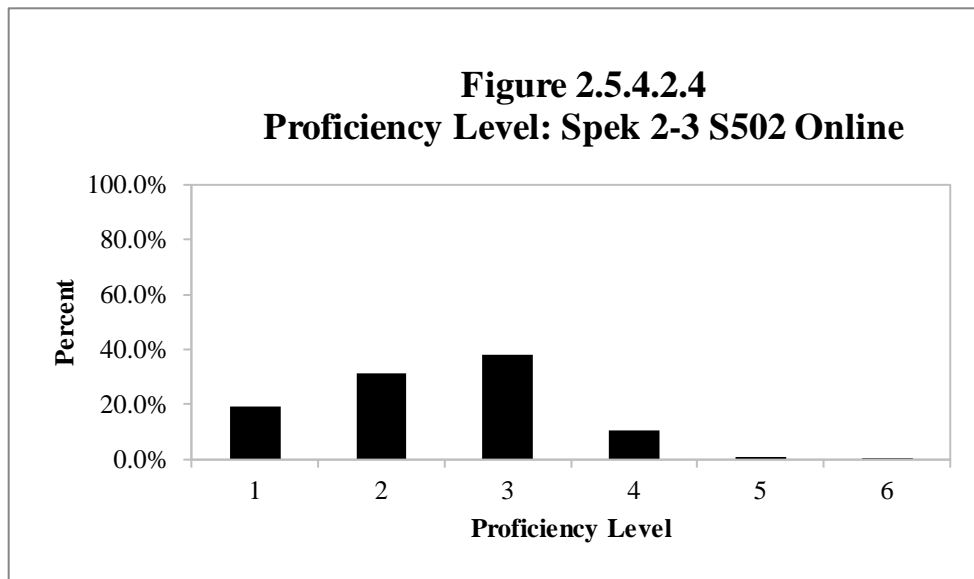


Table 2.5.4.2.4

Proficiency Level Distribution: Spek 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	22,491	18.50%	23,945	19.61%	46,436	19.06%
2	45,570	37.48%	30,463	24.95%	76,033	31.20%
3	41,370	34.03%	51,871	42.49%	93,241	38.27%
4	11,130	9.15%	14,738	12.07%	25,868	10.62%
5	905	0.74%	739	0.61%	1,644	0.67%
6	116	0.10%	328	0.27%	444	0.18%
Total	121,582	100.00%	122,084	100.00%	243,666	100.00%



2.5.4.3 Grades 4–5

Table 2.5.4.3.1

Proficiency Level Distribution: Spek 4-5 Pre-A S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	1,247	100.00%	1,990	100.00%	3,237	100.00%
Total	1,247	100.00%	1,990	100.00%	3,237	100.00%

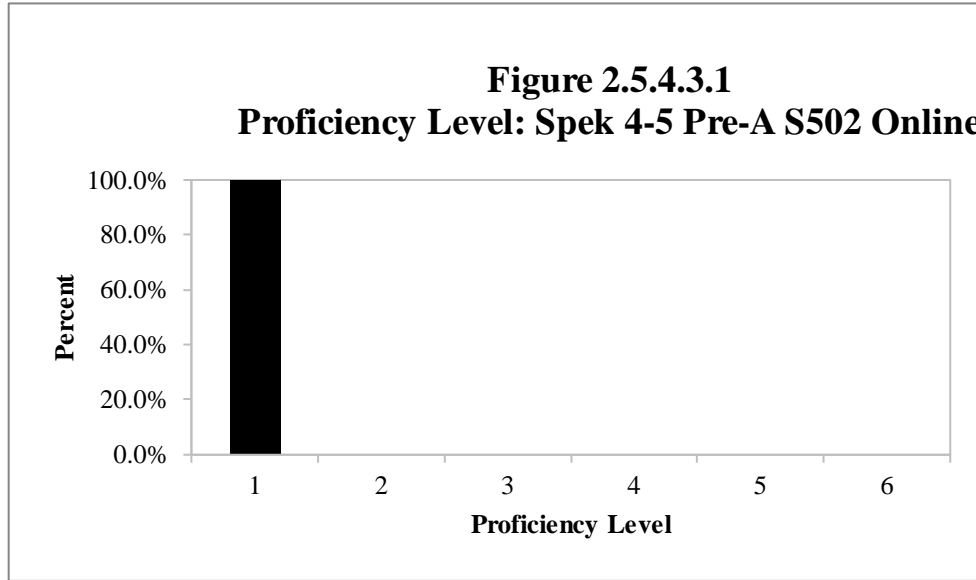


Table 2.5.4.3.2

Proficiency Level Distribution: Spek 4-5 A S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	6,948	42.87%	5,461	40.71%	12,409	41.89%
2	5,613	34.63%	4,716	35.15%	10,329	34.87%
3	3,115	19.22%	2,788	20.78%	5,903	19.93%
4	530	3.27%	446	3.32%	976	3.29%
5	2	0.01%	5	0.04%	7	0.02%
6	0	0.00%	0	0.00%	0	0.00%
Total	16,208	100.00%	13,416	100.00%	29,624	100.00%

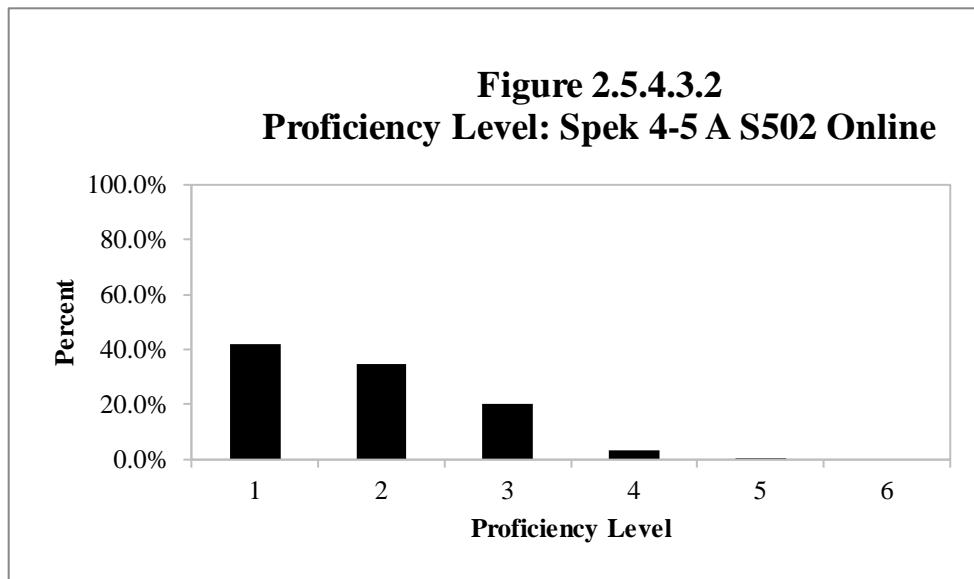


Table 2.5.4.3.3

Proficiency Level Distribution: Spek 4-5 B/C S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	6,146	5.67%	7,061	8.37%	13,207	6.86%
2	26,136	24.13%	26,510	31.43%	52,646	27.33%
3	48,055	44.37%	35,760	42.40%	83,815	43.51%
4	25,691	23.72%	14,191	16.83%	39,882	20.70%
5	2,157	1.99%	733	0.87%	2,890	1.50%
6	123	0.11%	88	0.10%	211	0.11%
Total	108,308	100.00%	84,343	100.00%	192,651	100.00%

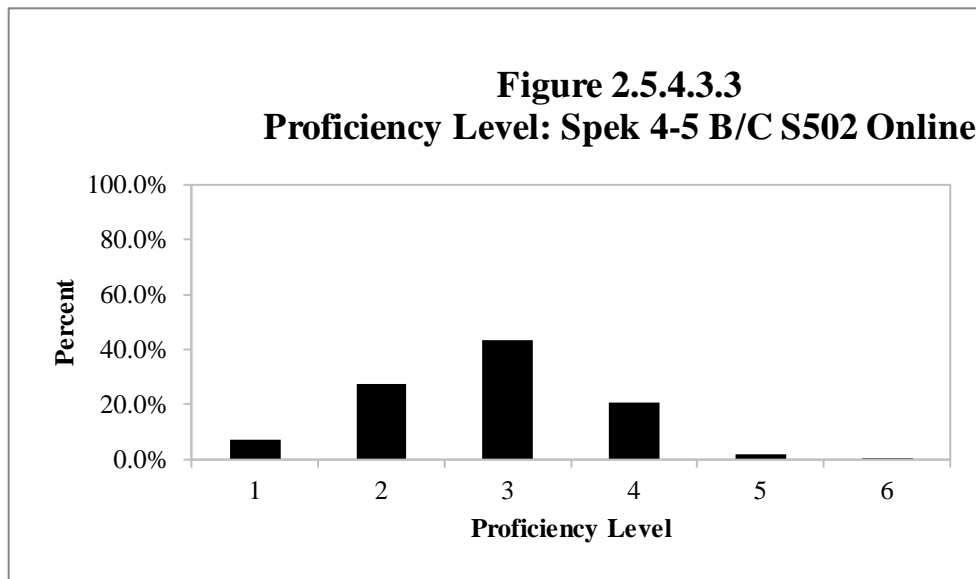
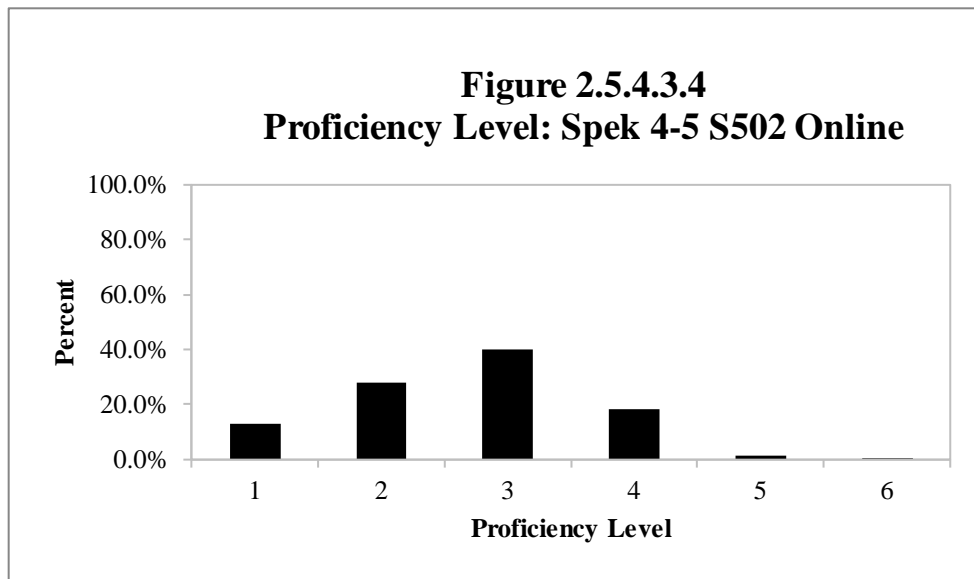


Table 2.5.4.3.4

Proficiency Level Distribution: Spek 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	14,341	11.40%	14,512	14.55%	28,853	12.79%
2	31,749	25.25%	31,226	31.30%	62,975	27.93%
3	51,170	40.69%	38,548	38.64%	89,718	39.78%
4	26,221	20.85%	14,637	14.67%	40,858	18.12%
5	2,159	1.72%	738	0.74%	2,897	1.28%
6	123	0.10%	88	0.09%	211	0.09%
Total	125,763	100.00%	99,749	100.00%	225,512	100.00%



2.5.4.4 Grades 6–8

Table 2.5.4.4.1

Proficiency Level Distribution: Spek 6-8 Pre-A S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	1,275	100.00%	1,661	100.00%	2,828	100.00%	5,764	100.00%
Total	1,275	100.00%	1,661	100.00%	2,828	100.00%	5,764	100.00%

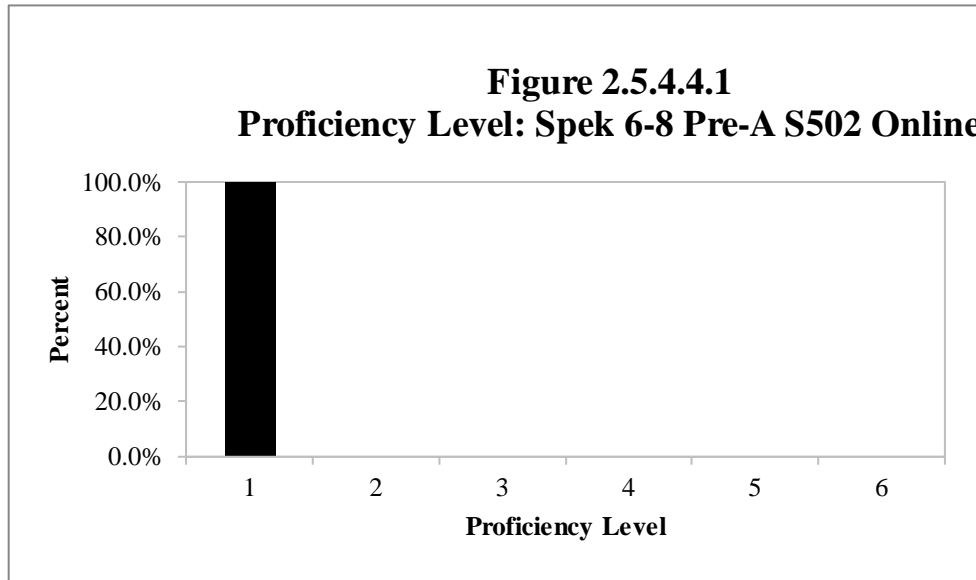


Table 2.5.4.4.2

Proficiency Level Distribution: Spek 6-8 A S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	6,213	47.37%	5,955	50.71%	10,318	47.23%	22,486	48.14%
2	4,979	37.96%	4,190	35.68%	5,415	24.79%	14,584	31.23%
3	1,725	13.15%	1,409	12.00%	5,813	26.61%	8,947	19.16%
4	198	1.51%	190	1.62%	292	1.34%	680	1.46%
5	0	0.00%	0	0.00%	9	0.04%	9	0.02%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	13,115	100.00%	11,744	100.00%	21,847	100.00%	46,706	100.00%

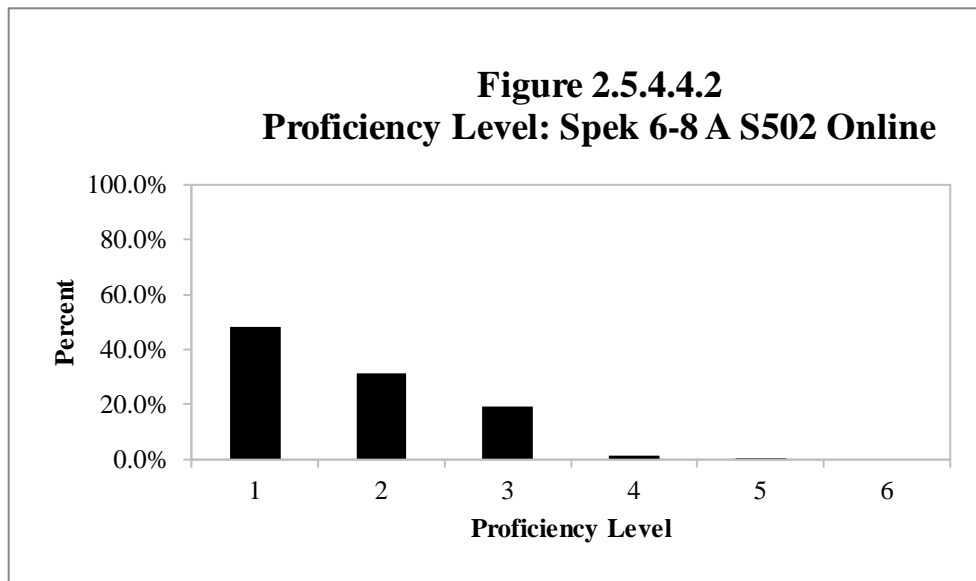


Table 2.5.4.4.3

Proficiency Level Distribution: Spek 6-8 B/C S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	5,830	8.72%	7,387	10.87%	4,969	10.26%	18,186	9.92%
2	19,529	29.21%	23,058	33.92%	11,569	23.88%	54,156	29.55%
3	29,978	44.83%	27,818	40.92%	25,500	52.63%	83,296	45.44%
4	11,361	16.99%	9,418	13.85%	6,265	12.93%	27,044	14.75%
5	158	0.24%	275	0.40%	114	0.24%	547	0.30%
6	10	0.01%	25	0.04%	32	0.07%	67	0.04%
Total	66,866	100.00%	67,981	100.00%	48,449	100.00%	183,296	100.00%

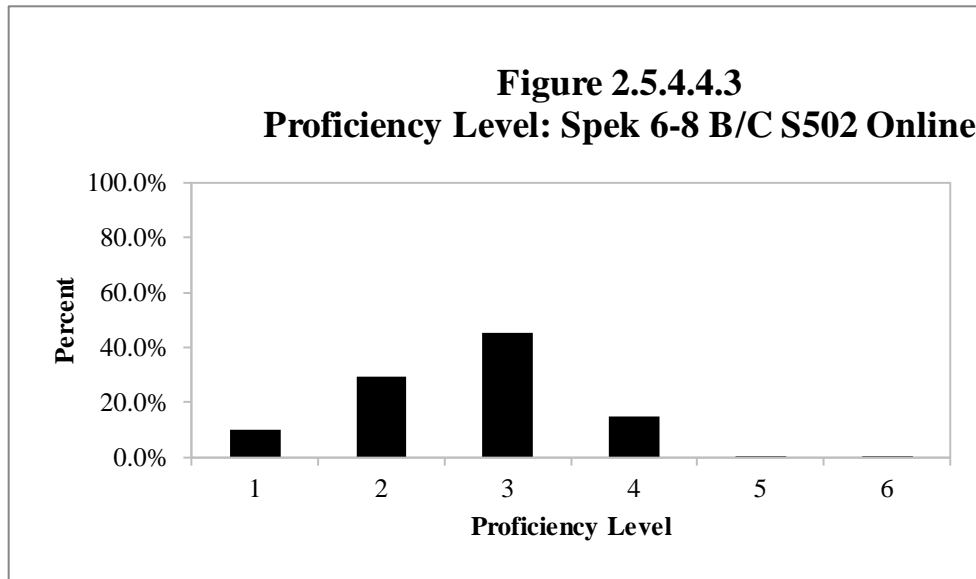
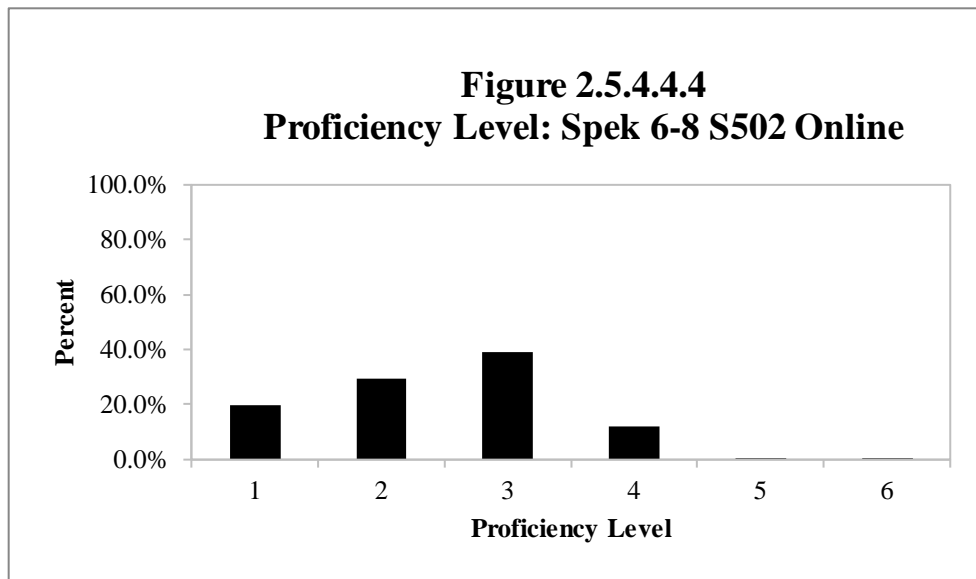


Table 2.5.4.4.4

Proficiency Level Distribution: Spek 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	13,318	16.39%	15,003	18.43%	18,115	24.77%	46,436	19.70%
2	24,508	30.16%	27,248	33.48%	16,984	23.23%	68,740	29.16%
3	31,703	39.02%	29,227	35.91%	31,313	42.82%	92,243	39.12%
4	11,559	14.23%	9,608	11.81%	6,557	8.97%	27,724	11.76%
5	158	0.19%	275	0.34%	123	0.17%	556	0.24%
6	10	0.01%	25	0.03%	32	0.04%	67	0.03%
Total	81,256	100.00%	81,386	100.00%	73,124	100.00%	235,766	100.00%



2.5.4.5 Grades 9–12

Table 2.5.4.5.1

Proficiency Level Distribution: Spek 9-12 Pre-A S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	2,366	100.00%	3,951	100.00%	3,125	100.00%	2,546	100.00%	11,988	100.00%
Total	2,366	100.00%	3,951	100.00%	3,125	100.00%	2,546	100.00%	11,988	100.00%

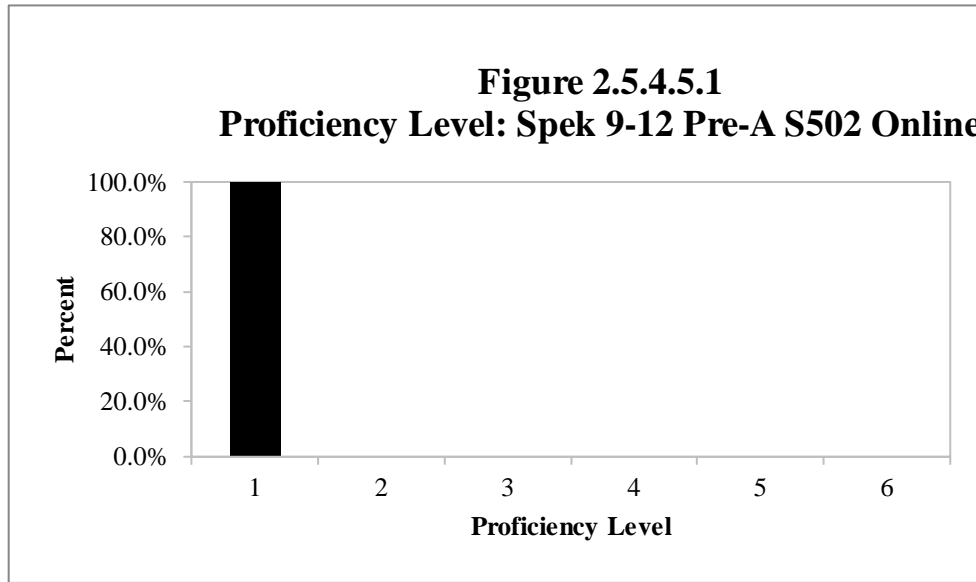


Table 2.5.4.5.2

Proficiency Level Distribution: Spek 9-12 A S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	11,987	39.81%	12,337	61.29%	5,289	66.22%	6,069	42.72%	35,682	49.26%
2	12,245	40.66%	4,397	21.84%	1,620	20.28%	3,775	26.58%	22,037	30.42%
3	5,140	17.07%	3,285	16.32%	1,039	13.01%	4,158	29.27%	13,622	18.81%
4	736	2.44%	111	0.55%	39	0.49%	203	1.43%	1,089	1.50%
5	5	0.02%	0	0.00%	0	0.00%	0	0.00%	5	0.01%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	30,113	100.00%	20,130	100.00%	7,987	100.00%	14,205	100.00%	72,435	100.00%

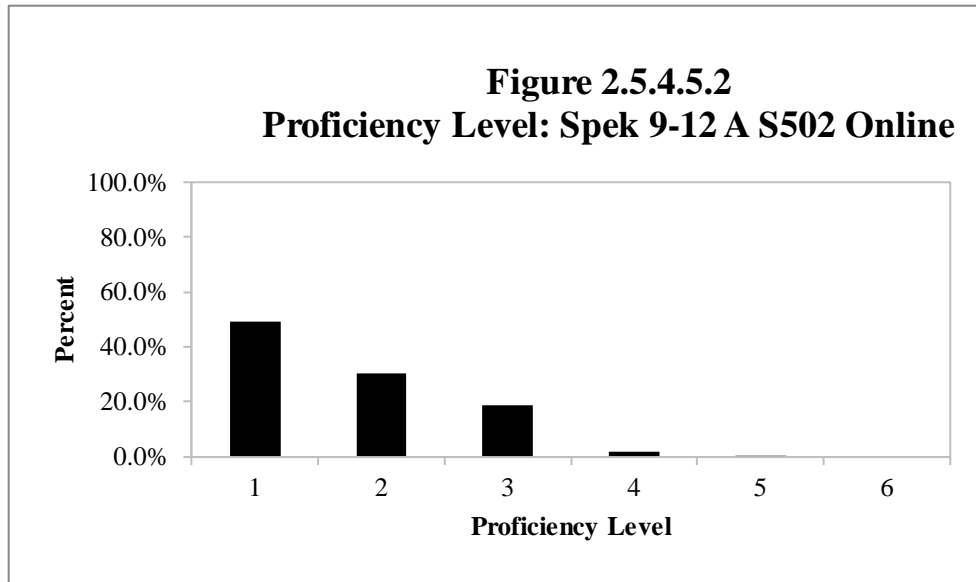


Table 2.5.4.5.3

Proficiency Level Distribution: Spek 9-12 B/C S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	2,262	7.74%	3,730	12.14%	4,936	14.86%	2,424	13.86%	13,352	12.06%
2	9,999	34.21%	8,904	28.97%	9,652	29.05%	5,852	33.45%	34,407	31.09%
3	13,346	45.66%	15,817	51.46%	17,189	51.74%	8,652	49.45%	55,004	49.69%
4	3,505	11.99%	2,248	7.31%	1,384	4.17%	517	2.96%	7,654	6.92%
5	108	0.37%	27	0.09%	47	0.14%	34	0.19%	216	0.20%
6	9	0.03%	11	0.04%	17	0.05%	16	0.09%	53	0.05%
Total	29,229	100.00%	30,737	100.00%	33,225	100.00%	17,495	100.00%	110,686	100.00%

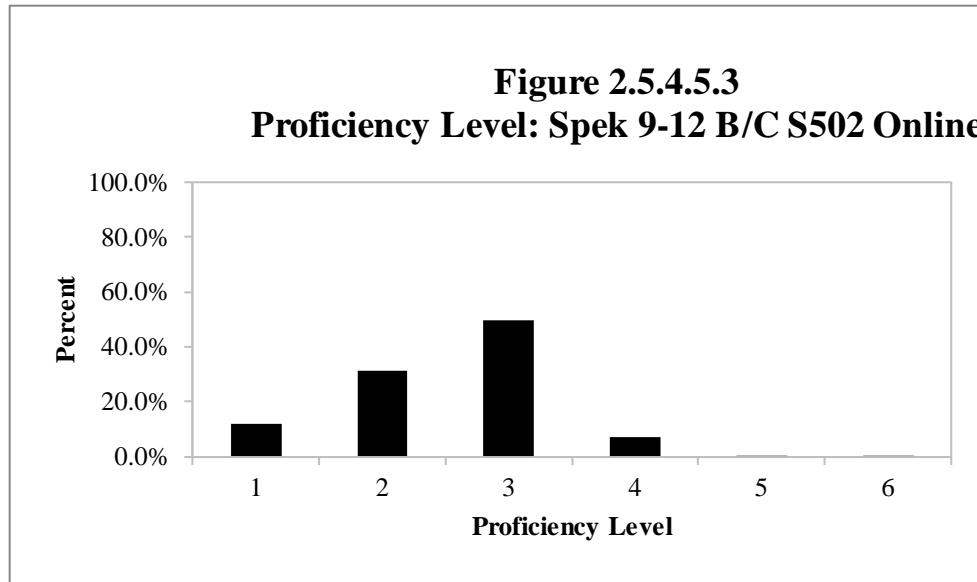
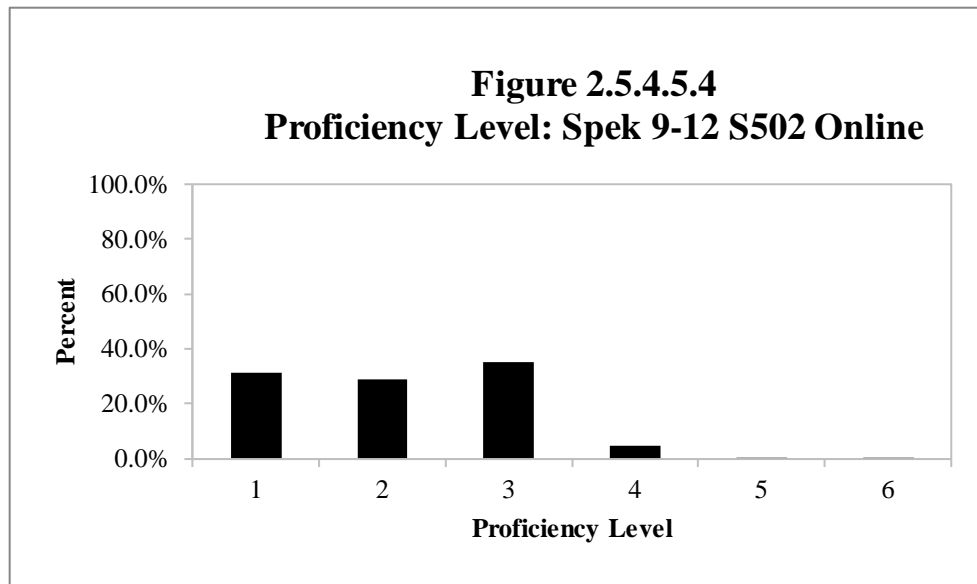


Table 2.5.4.5.4

Proficiency Level Distribution: Spek 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	16,615	26.93%	20,018	36.52%	13,350	30.11%	11,039	32.23%	61,022	31.28%
2	22,244	36.05%	13,301	24.26%	11,272	25.42%	9,627	28.11%	56,444	28.93%
3	18,486	29.96%	19,102	34.85%	18,228	41.11%	12,810	37.41%	68,626	35.17%
4	4,241	6.87%	2,359	4.30%	1,423	3.21%	720	2.10%	8,743	4.48%
5	113	0.18%	27	0.05%	47	0.11%	34	0.10%	221	0.11%
6	9	0.01%	11	0.02%	17	0.04%	16	0.05%	53	0.03%
Total	61,708	100.00%	54,818	100.00%	44,337	100.00%	34,246	100.00%	195,109	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 shows all possible raw scores. The following column shows the corresponding scale score. The next column shows the conditional standard error of measurement (CSEM) in the metric of the scale score, multiplied by 1.96. This is the confidence band as reported on students' score reports. For additional detail on standard error, 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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	111	256	1.0
1	199	45	1.6
2	213	32	1.8
3	222	28	1.8
4	230	27	1.9
5	237	28	1.9
6	246	31	2.2
7	257	35	2.5
8	270	39	2.8
9	286	41	3.1
10	304	42	3.4
11	321	42	3.7
12	338	40	4.0
13	354	38	4.3
14	367	36	4.6
15	380	36	4.9
16	394	40	5.5
17	414	52	6.0
18	445	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	111	256	1.0
1	213	45	1.8
2	226	32	1.9
3	236	28	1.9
4	243	27	2.1
5	251	28	2.3
6	260	31	2.5
7	270	35	2.8
8	284	39	3.1
9	300	41	3.4
10	317	42	3.6
11	335	42	3.9
12	352	40	4.3
13	367	38	4.6
14	381	36	4.9
15	394	36	5.5
16	408	40	6.0
17	428	52	6.0
18	459	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	133	233	1.0	1.0
1	202	45	1.6	1.6
2	216	33	1.7	1.7
3	225	29	1.8	1.8
4	233	28	1.9	1.8
5	241	28	1.9	1.9
6	249	31	2.1	2.0
7	260	35	2.4	2.3
8	273	39	2.8	2.7
9	289	41	3.1	3.0
10	307	42	3.4	3.3
11	325	42	3.7	3.6
12	342	40	4.0	3.9
13	357	38	4.3	4.2
14	370	36	4.6	4.5
15	383	36	4.8	4.7
16	398	40	5.4	5.1
17	417	52	6.0	5.9
18	449	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	133	252	1.0	1.0
1	206	45	1.7	1.6
2	220	33	1.8	1.7
3	229	29	1.8	1.8
4	237	28	1.9	1.9
5	245	28	2.0	1.9
6	254	31	2.3	2.1
7	264	35	2.5	2.4
8	277	39	2.9	2.8
9	293	41	3.2	3.1
10	311	42	3.5	3.4
11	329	42	3.8	3.7
12	346	40	4.1	4.0
13	361	38	4.4	4.3
14	374	36	4.7	4.5
15	387	36	4.9	4.8
16	402	40	5.6	5.3
17	422	52	6.0	6.0
18	453	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 S502 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	265	27	1.9	1.9
5	273	28	2.3	2.2
6	282	31	2.7	2.5
7	292	35	3.0	2.9
8	306	39	3.2	3.2
9	322	41	3.5	3.4
10	339	42	3.8	3.7
11	357	42	4.1	4.0
12	374	40	4.4	4.3
13	389	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	481	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	155	256	1.0	1.0
1	274	45	2.3	2.2
2	288	32	3.0	2.8
3	296	28	3.1	3.0
4	304	27	3.2	3.1
5	312	28	3.3	3.3
6	320	31	3.5	3.4
7	331	35	3.6	3.6
8	344	39	3.8	3.8
9	361	41	4.2	4.0
10	378	42	4.5	4.4
11	396	42	4.9	4.7
12	413	40	5.5	5.2
13	428	38	6.0	5.8
14	442	36	6.0	6.0
15	455	36	6.0	6.0
16	469	40	6.0	6.0
17	488	52	6.0	6.0
18	520	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	188	133	1.2	1.1	1.0
1	231	45	1.6	1.5	1.4
2	244	32	1.7	1.6	1.6
3	253	28	1.8	1.7	1.6
4	261	27	1.9	1.8	1.7
5	269	28	2.0	1.9	1.8
6	277	31	2.3	2.1	1.9
7	288	35	2.6	2.4	2.2
8	301	39	3.0	2.8	2.6
9	317	41	3.3	3.1	3.0
10	335	42	3.5	3.4	3.3
11	353	42	3.8	3.7	3.6
12	370	40	4.1	4.0	3.9
13	385	38	4.4	4.3	4.2
14	398	36	4.7	4.5	4.5
15	411	36	4.9	4.8	4.7
16	425	40	5.4	5.1	5.0
17	445	52	6.0	5.8	5.6
18	477	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	188	185	1.2	1.1	1.0
1	246	45	1.7	1.7	1.6
2	260	32	1.9	1.8	1.7
3	269	28	2.0	1.9	1.8
4	277	27	2.3	2.1	1.9
5	284	28	2.5	2.3	2.1
6	293	31	2.8	2.6	2.4
7	303	35	3.0	2.9	2.7
8	317	39	3.3	3.1	3.0
9	333	41	3.5	3.4	3.3
10	350	42	3.8	3.7	3.6
11	368	42	4.1	4.0	3.9
12	385	40	4.4	4.3	4.2
13	400	38	4.7	4.6	4.5
14	414	36	5.0	4.9	4.8
15	427	36	5.5	5.2	5.0
16	441	40	6.0	5.7	5.4
17	461	52	6.0	6.0	6.0
18	492	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	232	92	1.3	1.2	1.1	1.0
1	259	45	1.6	1.5	1.4	1.3
2	273	33	1.8	1.7	1.5	1.4
3	283	29	1.9	1.8	1.7	1.5
4	291	28	2.0	1.9	1.8	1.6
5	299	29	2.3	2.0	1.8	1.7
6	307	31	2.6	2.3	1.9	1.8
7	318	34	2.9	2.7	2.3	2.0
8	331	38	3.2	3.0	2.8	2.5
9	347	41	3.4	3.3	3.2	3.0
10	365	42	3.7	3.6	3.5	3.3
11	382	42	4.0	3.9	3.8	3.7
12	399	40	4.4	4.2	4.1	4.0
13	414	38	4.6	4.5	4.4	4.3
14	428	36	4.9	4.8	4.7	4.6
15	441	37	5.2	5.1	5.0	4.8
16	456	40	5.6	5.4	5.3	5.1
17	476	52	6.0	5.9	5.7	5.5
18	507	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	232	89	1.3	1.2	1.1	1.0
1	258	45	1.6	1.5	1.4	1.3
2	272	33	1.8	1.7	1.5	1.4
3	282	29	1.9	1.8	1.7	1.5
4	290	28	2.0	1.9	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	318	34	2.9	2.7	2.3	2.0
8	331	38	3.2	3.0	2.8	2.5
9	347	41	3.4	3.3	3.2	3.0
10	364	42	3.7	3.6	3.5	3.3
11	382	42	4.0	3.9	3.8	3.7
12	398	40	4.3	4.2	4.1	4.0
13	414	38	4.6	4.5	4.4	4.3
14	427	36	4.9	4.8	4.7	4.5
15	441	37	5.2	5.1	5.0	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	507	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	106	52	1.0
1	106	52	1.0
2	124	40	1.1
3	137	37	1.3
4	150	40	1.4
5	163	48	1.5
6	176	61	1.7

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	106	47	1.0
1	106	47	1.0
2	117	38	1.1
3	129	34	1.2
4	139	32	1.3
5	148	32	1.4
6	158	33	1.5
7	168	34	1.6
8	179	36	1.7
9	192	38	1.8
10	206	42	2.0
11	224	48	2.3
12	249	54	2.7
13	275	52	3.2
14	297	48	3.7
15	317	47	4.1
16	338	50	4.5
17	359	59	4.9
18	380	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
6	106	43	1.0
7	156	28	1.5
8	164	28	1.5
9	171	28	1.6
10	178	28	1.7
11	186	28	1.8
12	193	29	1.8
13	201	29	1.9
14	209	31	2.0
15	218	32	2.2
16	228	34	2.4
17	240	37	2.6
18	253	38	2.8
19	267	38	3.1
20	279	37	3.3
21	291	35	3.6
22	302	34	3.8
23	312	33	4.0
24	322	33	4.2
25	332	34	4.4
26	343	35	4.6
27	355	39	4.8
28	367	43	5.1
29	379	49	5.4
30	403	68	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	118	40	1.0	1.0
1	118	40	1.0	1.0
2	118	40	1.0	1.0
3	130	37	1.1	1.1
4	144	40	1.2	1.2
5	158	49	1.4	1.3
6	172	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 S502 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	124	34	1.1	1.0
4	134	33	1.2	1.1
5	144	33	1.2	1.2
6	154	34	1.3	1.3
7	165	36	1.4	1.4
8	178	37	1.6	1.5
9	191	39	1.7	1.6
10	206	42	1.8	1.7
11	225	48	2.0	1.9
12	249	55	2.5	2.3
13	275	52	3.0	2.8
14	298	48	3.5	3.3
15	318	47	3.9	3.7
16	338	50	4.3	4.1
17	358	58	4.6	4.4
18	378	73	5.0	4.8

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
6	118	34	1.0	1.0
7	156	30	1.4	1.3
8	164	30	1.4	1.3
9	173	30	1.5	1.4
10	181	30	1.6	1.5
11	189	30	1.7	1.6
12	197	30	1.7	1.6
13	205	30	1.8	1.7
14	214	31	1.9	1.8
15	223	32	2.0	1.9
16	233	34	2.2	1.9
17	245	37	2.4	2.2
18	258	38	2.7	2.4
19	271	38	2.9	2.7
20	284	37	3.2	3.0
21	295	35	3.4	3.2
22	306	34	3.6	3.4
23	317	33	3.8	3.6
24	327	33	4.0	3.8
25	337	34	4.2	4.0
26	348	36	4.5	4.2
27	361	39	4.7	4.5
28	374	44	5.0	4.7
29	387	51	5.3	5.0
30	425	87	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	130	55	1.0	1.0
1	133	52	1.0	1.0
2	151	40	1.2	1.1
3	164	37	1.3	1.2
4	177	40	1.4	1.3
5	190	47	1.5	1.4
6	203	61	1.6	1.5

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	130	51	1.0	1.0
1	130	51	1.0	1.0
2	146	39	1.1	1.1
3	158	34	1.2	1.2
4	168	33	1.3	1.2
5	178	33	1.4	1.3
6	188	35	1.5	1.4
7	200	36	1.6	1.5
8	212	38	1.7	1.6
9	226	39	1.8	1.7
10	241	42	1.9	1.8
11	259	48	2.2	2.0
12	284	55	2.8	2.5
13	310	52	3.3	3.1
14	332	48	3.7	3.6
15	352	47	4.1	4.0
16	373	50	4.5	4.4
17	394	59	4.9	4.7
18	415	75	5.4	5.2

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
6	130	49	1.0	1.0
7	192	31	1.5	1.4
8	201	31	1.6	1.5
9	210	31	1.7	1.6
10	219	30	1.7	1.6
11	227	30	1.8	1.7
12	236	30	1.9	1.8
13	244	30	1.9	1.8
14	253	31	2.1	1.9
15	262	33	2.3	2.0
16	272	34	2.5	2.3
17	284	36	2.8	2.5
18	297	38	3.0	2.8
19	310	38	3.3	3.1
20	322	37	3.5	3.4
21	334	35	3.8	3.6
22	345	34	4.0	3.8
23	356	33	4.2	4.1
24	366	33	4.4	4.2
25	376	34	4.6	4.4
26	387	36	4.8	4.6
27	400	39	5.0	4.8
28	413	44	5.4	5.1
29	426	51	5.7	5.5
30	443	63	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	148	48	1.0	1.0	1.0
1	148	48	1.0	1.0	1.0
2	162	40	1.1	1.1	1.1
3	175	37	1.2	1.2	1.1
4	188	40	1.3	1.3	1.2
5	201	48	1.4	1.4	1.3
6	214	61	1.5	1.5	1.4

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	148	45	1.0	1.0	1.0
1	148	45	1.0	1.0	1.0
2	158	39	1.1	1.1	1.0
3	170	35	1.2	1.2	1.1
4	181	34	1.3	1.2	1.2
5	192	35	1.4	1.3	1.3
6	204	37	1.5	1.4	1.4
7	217	39	1.6	1.5	1.5
8	231	40	1.7	1.6	1.6
9	246	40	1.8	1.7	1.7
10	261	43	1.9	1.8	1.8
11	280	49	2.2	2.0	1.9
12	305	55	2.8	2.7	2.5
13	332	52	3.4	3.2	3.1
14	354	48	3.8	3.7	3.5
15	374	47	4.2	4.0	3.9
16	395	50	4.6	4.4	4.3
17	416	59	4.9	4.8	4.6
18	437	76	5.5	5.3	5.1

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
6	148	44	1.0	1.0	1.0
7	208	33	1.5	1.4	1.4
8	218	33	1.6	1.5	1.5
9	227	32	1.6	1.6	1.5
10	236	31	1.7	1.6	1.6
11	245	30	1.8	1.7	1.7
12	253	30	1.8	1.8	1.7
13	261	30	1.9	1.8	1.8
14	270	31	2.0	1.9	1.8
15	279	32	2.2	2.0	1.9
16	289	35	2.5	2.3	2.1
17	301	37	2.7	2.6	2.4
18	314	39	3.0	2.9	2.7
19	327	38	3.3	3.1	3.0
20	340	37	3.6	3.4	3.3
21	352	35	3.8	3.6	3.5
22	363	34	4.0	3.8	3.7
23	373	33	4.2	4.0	3.9
24	383	33	4.4	4.2	4.1
25	393	34	4.5	4.4	4.2
26	404	35	4.7	4.6	4.4
27	416	38	4.9	4.8	4.6
28	428	43	5.3	5.0	4.9
29	440	49	5.6	5.4	5.2
30	463	67	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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	172	40	1.1	1.0	1.0	1.0
1	172	40	1.1	1.0	1.0	1.0
2	172	40	1.1	1.0	1.0	1.0
3	185	37	1.2	1.1	1.1	1.1
4	198	40	1.3	1.2	1.2	1.2
5	211	47	1.4	1.3	1.3	1.3
6	224	60	1.5	1.4	1.4	1.4

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	172	37	1.1	1.0	1.0	1.0
1	172	37	1.1	1.0	1.0	1.0
2	172	37	1.1	1.0	1.0	1.0
3	181	35	1.1	1.1	1.1	1.0
4	192	34	1.2	1.2	1.1	1.1
5	203	36	1.3	1.3	1.2	1.2
6	215	38	1.4	1.4	1.3	1.3
7	229	40	1.5	1.5	1.4	1.4
8	244	40	1.6	1.6	1.5	1.5
9	258	40	1.7	1.7	1.6	1.6
10	274	43	1.8	1.8	1.8	1.7
11	292	49	2.0	1.9	1.9	1.9
12	317	55	2.7	2.5	2.4	2.3
13	344	52	3.2	3.1	3.1	3.0
14	367	47	3.6	3.5	3.4	3.4
15	386	46	4.0	3.8	3.7	3.6
16	407	50	4.4	4.2	4.1	4.0
17	428	59	4.7	4.6	4.5	4.4
18	449	76	5.3	5.1	4.9	4.8

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 S502 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
6	172	36	1.1	1.0	1.0	1.0
7	219	34	1.4	1.4	1.3	1.3
8	229	33	1.5	1.5	1.4	1.4
9	239	32	1.6	1.5	1.5	1.5
10	248	31	1.6	1.6	1.6	1.5
11	256	30	1.7	1.7	1.6	1.6
12	265	30	1.8	1.7	1.7	1.7
13	273	30	1.8	1.8	1.8	1.7
14	281	31	1.9	1.8	1.8	1.8
15	291	33	2.0	1.9	1.9	1.9
16	301	35	2.2	2.1	2.0	1.9
17	313	37	2.6	2.4	2.3	2.2
18	326	38	2.9	2.8	2.7	2.6
19	339	38	3.1	3.1	3.0	2.9
20	352	37	3.4	3.3	3.2	3.1
21	364	35	3.6	3.5	3.4	3.3
22	375	34	3.8	3.7	3.6	3.5
23	385	33	4.0	3.8	3.7	3.6
24	395	33	4.1	4.0	3.9	3.8
25	405	34	4.3	4.2	4.0	3.9
26	416	36	4.5	4.4	4.3	4.2
27	429	39	4.8	4.6	4.5	4.4
28	442	44	5.0	4.9	4.8	4.7
29	455	51	5.5	5.3	5.1	5.0
30	476	67	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 on each Online ACCESS for ELLs test form are new, as determined by the refreshment plan for that series. For Series 502, we refreshed all four domains. For Listening and Reading domains, WIDA implements a multi-year 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 2018, WIDA and CAL assessment experts reviewed the 501 Listening and Reading item pools and identified folders that they believed the team should refresh for Series 502, according to the targeted refreshment plan. To meet these Series 502 targets, seven Listening folders were field tested and between eight and 13 Reading folders were field tested by grade clusters. For the Writing and Speaking domains, which are shorter and performance based, and which have additional content and exposure considerations in terms of task refreshment, the refreshment plan was determined by WIDA and CAL assessment experts 3 years ahead to ensure that the test development effort can accommodate the refreshment target set for each series. Since we need to anchor one task to the Writing scales, we can only refresh one of the two Writing tasks. At the same time, Speaking has three panels, each with three unique tasks, so we refreshed one panel, or three tasks, for Series 502.

We used an equating procedure known as common item equating to equate the results on the new item pool and forms to the older item pool and forms using the common items, which are items that appear in both Series 501 and 502. The characteristics of the common items are kept the same between series, as are the wording, formatting, and other test characteristics such as graphics. Furthermore, common items appear in approximately the same item sequence position as they appeared in the previous test series. We designed all common items to be initial anchors unless we later removed them from the anchor set. We kept common items that displayed adequate item fit and exhibited no C-level DIF. We did not need to remove any common items from the anchor set for these concerns. In this procedure, we keep the difficulty measures for items that appear on both the new and the old item pools and test forms constant across both item pools and forms. In this way, the reader may use the newer form to interpret the students' scores as reflective of their performance, by using the same frame of reference.

We used a pre-equating design to conduct the annual equating for Listening and Reading. This design allowed for Listening and Reading item parameters to be available for setting up the computer adaptive engine prior to operational administration. For the Listening and Reading domains, we used student data collected from the Series 502 embedded field test to conduct the equating analyses. The interruption in school and testing, due to the COVID-19 pandemic, did not affect the annual equating analysis for the Listening and Reading domains, as we collected the pre-equating data before the school closures in spring 2020. We included in the analyses all the student data that was available at the time that we conducted the equating analyses.

Both Writing and Speaking used an appended field test design. Preliminary task parameters had been obtained using student data collected during the Series 502 field test administration in spring 2020, and those parameters were to be verified using Series 502 operational data collected in spring 2021. Due to the schooling and testing interruptions caused by the COVID-19 pandemic, we implemented a pre-equating procedure instead. Because we collected the Series 502 field test data in the spring of 2020, before the school closures, we believe these data might be less affected by COVID-19 as compared to the Series 502 operational data, which we collected in the spring of 2021, and therefore we used the Writing and Speaking Series 502 appended field test data to conduct annual equating.

For Writing, DRC drew random samples of students among all available student data at the time of the Writing data draw, according to WIDA's predetermined sampling plan. In the 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 Speaking, DRC drew random samples of students among all available student data at the time of the Speaking data draw. We fixed the number of students to 500 per field test task, and we included in the analysis all the student data that was available at the time that we conducted the equating analyses.

For the Listening and Reading domains, because we include all Series 501 operational items in the initial anchor set in conducting the annual equating, the content representation of the anchor set is not a concern.

For the Speaking and Writing domains, it is important to note the overall assessment construct when we consider the distribution of anchor tasks. The overarching goal of ACCESS for ELLs Online is to measure academic English language proficiency of students. Proficiency is measured according to a proficiency 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). Because of this conceptualization of the WIDA ELD Standards, scores are not reported for each of the Standards, and it is not necessary to assess all five Standards in one domain, as long as each of the Standards is measured on the assessment in some capacity (although ACCESS for ELLs Online does strive to represent all five WIDA Standards in each domain subtest). Therefore, it is not necessary for the anchor set to contain tasks in all five of the WIDA Standards. Rather, it is more important to ensure that each task assesses the targeted PLs so we can sufficiently claim that ACCESS for ELLs Online truly measures across the breadth of the proficiency scale. Thus, the set of anchor tasks for the Writing and Speaking subtests need to assess across the breadth of proficiency levels in the same way the entire test does; the unanchored tasks similarly are designed to assess the breadth of the proficiency scale.

The Writing domain consists of two sets of operational tasks: the first targeting the Language of Language Arts and the Language of Social Studies and the second targeting the Language of Mathematics and the Language of Science. Each set assesses the entire breadth of the proficiency scale, from PL 1 to 5. One of the two sets is refreshed in a given year on an alternating schedule,

so the WIDA Standard of the anchor task and the unanchored task alternates each year. The sets of anchor tasks target across the proficiency scale, mirroring the targeted proficiency levels of the Writing subtest as a whole, which supports the argument that ACCESS for ELLs Online assesses academic English language proficiency in the domain of Writing.

The Speaking domain consists of three sets of operational tasks: the first targeting Social and Instructional Language; the second targeting the Language of Language Arts and the Language of Social Studies; and the third targeting the Language of Mathematics and the Language of Science. Each set assesses the entire breadth of the proficiency scale, from PL 1 to 5. Generally, one or two of the three sets are refreshed in a given year on a rotating schedule, so the WIDA Standard of the anchor task also rotates. The anchor task will always measure the breadth of the proficiency scale, which is crucial to the construct of academic English language proficiency in the domain of Speaking. The anchor task samples the WIDA ELD Standards in the same way that the unanchored tasks do, so overall, the Speaking domain samples all five of the WIDA ELD Standards. This allows for the Speaking domain to be of manageable length and still contain embedded field test tasks, given the seat time required of students to complete each Speaking performance task.

We anchored all items common to both Series 501 and 502 item pools and forms to their 501 values in the first equating run. After the first equating run, some items that we had originally anchored, either to their operational or to their field test value, proved to have changed in their levels of difficulty. The “displacement” statistic is a measure of this change. This statistic shows the difference between the difficulty value of the anchored item and what its difficulty value would have been had we not anchored it. 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 for Writing tasks and PL 3 and 5 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 PL 1 tasks, we used a slightly different displacement criterion (above 0.50 or below -0.50) since anchored PL 1 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 PL 1 tasks to be very easy, and therefore we can expect most students (98% to 99%) to get the full 2 points. As a result, the item difficulties for these PL 1 tasks are susceptible to small sampling fluctuations. A slight change in the percentages of students getting the full 2 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 PL 1 tasks would need to be re-estimated each time a slightly different sample was used to estimate them. Therefore, we used a more conservative estimate (-0.5 to 0.5) to evaluate the displacement statistics for the Speaking PL 1 tasks in order to ensure the stability of the Speaking scale scores.

The tables that follow present a summary of the equating procedures. The first section of each table compares the current test (i.e., the Series 502 version of that item pool and test form) to the

previous year's test (i.e., the Series 501 version of that item pool and test form). The table shows the number of items, the average item difficulty, the standard deviation of the item difficulty values, and the difficulty value of the easiest and hardest item 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 were 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 a large number of test items in the item pool, as well as large anchor sets. Additionally, the Series 502 Writing domain tests consist of only two tasks, with only one task serving as an anchor between series. Similarly, we might expect some differences in the average difficulty values for the two Speaking series, as the test forms include only nine tasks, and only two-thirds of the test serves as the anchor between series.

The second section of each table presents information about the anchoring items (or tasks) and shows the total number of possible anchors that we initially anchored to the values of the previous series as well as the average item (or task) difficulty and the average standard deviation of the difficulty values for those items (or tasks). Next, the table shows the number of items (or tasks) that we anchored in the final equating run, again with the average item (or task) difficulty and the average standard deviation of those difficulty values for those items (or tasks). Finally, the table gives the percentage of items (or tasks) that served as anchors and their average displacement values. In general, the larger the number and the higher the percentage of items (or tasks) anchored and the closer their average displacement is to 0.00, the more trustworthy the equating results will be (Johns & Smith, 2006; Stahl & Timothy Muckle, 2007). For the Listening and Reading domains, we expected the average displacements to be around 0.00 since there were high percentages of items anchored. For the Writing domain, when there was only one task anchored to the known value derived from the special research study, Winsteps automatically set the displacement statistic for the anchor to 0, and the average displacement statistic was also 0.

The third section of each table gives information about the anchor items (or tasks), both by order of displacement statistics and by order of item difficulty. The displacement statistics provide information regarding the difference between the difficulty value of each anchored item (or task) and what that difficulty value would have been had we not anchored the item (or task). Smaller displacement statistics indicate more consistency between the item's (or task's) difficulty value on the Series 501 test form and on the Series 403 test form. Lastly, it is desirable that the anchor items on an item pool and test form represent a similar range of difficulties as the entire item pool and test form (Kolen & Brennan, 2004).

For the Writing and Speaking tasks, these tables 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 ability 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 scale categories. That is, a step measure indicates how likely it is for a student to receive a score in a particular rating scale category relative to the adjacent 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 rating 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 in order to be able to make valid inferences regarding a student's level of ability based on his/her 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 rating scale categories in a similar manner when assigning scores to students' responses to all tasks included on the test form. That is, under this assumption, when Winsteps analyzed the students' 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 to the other tasks. Similarly, the computer program treated the 4s that raters assigned to students' responses to a task as equivalent to the 4s that raters assigned to students' responses to the other tasks, and so on. Accordingly, the output from the Winsteps analysis reported a single set of step measures that applied to all 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 7 to Step 8). While these findings do not signal optimal rating 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 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 first three score categories on a five-category (0–4) rating 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 all five score categories on that rating 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. The advances in the step measures ranged from 0.34 logits (from Step 2 to Step 3) to 1.52 logits (from Step 3 to Step 4). While these findings do not signal optimal rating scale functioning (i.e., the step measures did not advance from step to step by at least 1.4 logits), raters' use of the Speaking scale 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 502 Listening and Reading test forms were similar to those for the previous series for all grade-level clusters. 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 average difficulty levels for the items appearing on the Writing test forms were similar to those for the previous series for all grade-level clusters and tiers, except for Grade 1 Tier A, Grade 1 Tier B/C, and Grades 2–3 Tier B/C.

The average difficulty levels for the tasks appearing on the Series 501 Speaking test forms were similar to those for the previous series for all grade-level clusters. For each Speaking test form, the anchor tasks represented a range of difficulties that spanned nearly the entire task difficulty continuum.

For the Listening domain, the percentages of items anchored in the final equating runs ranged from 67% to 85%, and the average displacement statistics were either equal to or close to 0.00. None of the displacement statistics were above 0.30 or below -0.30, as to be expected.

For the Reading domain, the percentages of items anchored in the final equating runs ranged from 58% to 71%, and the average displacement statistics were either equal to or close to 0.00. None of the displacement statistics were above 0.30 or below -0.30, as to be expected.

For the Writing domain, the displacement statistic for the anchor task was automatically set to 0 in Winsteps; therefore, the average displacement statistic was also 0.

For the Speaking domain, the percentages of tasks anchored in the final equating runs were between 22% and 33%, and the average displacement statistics were all close to 0.00. None of the displacement statistics were above 0.50 or below -0.50, as to be expected.

WIDA Psychometricians reviewed the equating plans before CAL conducted the equating analyses. The Psychometricians then reviewed the equating results at the conclusion of the equating project to ensure that the equating was done correctly and the results were deemed reasonable. In addition, WIDA and CAL Psychometricians reviewed the annual equating results and identified issues that they felt they needed to bring up to the WIDA Technical Advisory Committee.

2.7.1 Listening

2.7.1.1 Grade 1

Table 2.7.1.1

Equating Summary: List 1 S502 Online

Comparison of Forms	Form 502		Form 501			
	No. of Items	Average Difficulty (Std. Dev.)	No. of Items	Average Difficulty (Std. Dev.)		
	54	-1.06 (1.15)	54	-1.20 (1.28)		
	Easiest	Hardest	Easiest	Hardest		
-3.59	1.34	-4.38	1.34			
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	45	-1.01 (1.12)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	40	-0.87 (1.09)				
	Percentage Anchors	Average Displacement				
74%	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
	13878	0.31	-0.25	14952	-3.03	0.01
	13877	1.34	-0.24	17781	-2.96	-0.19
	17814	-1.12	-0.23	13891	-2.55	0.16
	13879	-0.64	-0.21	17813	-2.32	-0.02
	17781	-2.96	-0.19	13890	-2.23	0.24
	11667	0.44	-0.19	13816	-2.22	0.06
	12846	0.58	-0.11	11671	-2.02	-0.02
	14898	-1.93	-0.10	14898	-1.93	-0.10
	14897	-1.30	-0.09	16531	-1.79	-0.08
	16531	-1.79	-0.08	13900	-1.63	0.07
	16533	-0.47	-0.04	17791	-1.57	0.14
	11671	-2.02	-0.02	14899	-1.50	0.04
	13898	-0.94	-0.02	14953	-1.43	0.12
	12848	0.10	-0.02	14897	-1.30	-0.09
17813	-2.32	-0.02	16648	-1.24	0.21	

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	16535	-1.22	0.00	16535	-1.22	0.00
14952	-3.03	0.01	13899	-1.15	0.07	
16649	1.13	0.01	13814	-1.12	0.03	
13814	-1.12	0.03	17814	-1.12	-0.23	
12847	0.02	0.03	13898	-0.94	-0.02	
16560	-0.02	0.03	16641	-0.86	0.05	
14899	-1.50	0.04	11668	-0.81	0.05	
11668	-0.81	0.05	16642	-0.74	0.08	
16641	-0.86	0.05	13879	-0.64	-0.21	
13815	-0.38	0.06	16533	-0.47	-0.04	
13816	-2.22	0.06	13815	-0.38	0.06	
13899	-1.15	0.07	17793	-0.27	0.08	
13900	-1.63	0.07	16640	-0.25	0.14	
16558	-0.15	0.07	16558	-0.15	0.07	
16642	-0.74	0.08	16560	-0.02	0.03	
17793	-0.27	0.08	17788	0.01	0.11	
17788	0.01	0.11	12847	0.02	0.03	
14953	-1.43	0.12	12848	0.10	-0.02	
16640	-0.25	0.14	13878	0.31	-0.25	
17791	-1.57	0.14	11667	0.44	-0.19	
16559	0.50	0.16	16559	0.50	0.16	
13891	-2.55	0.16	12846	0.58	-0.11	
16648	-1.24	0.21	16650	0.63	0.28	
13890	-2.23	0.24	16649	1.13	0.01	
16650	0.63	0.28	13877	1.34	-0.24	

2.7.1.2 Grades 2–3

Table 2.7.1.2

Equating Summary: List 2-3 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	54	-1.15 (1.62)		54	-1.20 (1.60)	
	Easiest	Hardest		Easiest	Hardest	
-4.25	2.02		-4.25	2.02		
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	48	-1.12 (1.64)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	46	-1.05 (1.61)				
	Percentage Anchors	Average Displacement				
85%	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
	13789	-0.72	-0.28	17769	-4.25	0.06
	13790	-2.68	-0.25	14879	-4.12	0.21
	11546	-1.50	-0.21	11544	-3.56	0.09
	12828	-2.29	-0.16	13905	-3.26	0.14
	12734	0.26	-0.15	12825	-3.26	-0.11
	16602	-2.56	-0.12	13904	-3.24	0.09
	12706	-1.64	-0.12	12956	-3.09	0.15
	12786	1.56	-0.11	17770	-2.72	0.18
	11545	-1.90	-0.11	13790	-2.68	-0.25
	12825	-3.26	-0.11	14884	-2.64	0.05
	12988	-1.07	-0.10	16602	-2.56	-0.12
	12830	-1.17	-0.10	13910	-2.33	0.05
	12705	-0.17	-0.04	12828	-2.29	-0.16
	16654	1.15	-0.04	13906	-2.24	0.02
12785	0.00	-0.04	12954	-1.97	-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
	16685	0.53	-0.03	11545	-1.90	-0.11
16603	-0.35	-0.03	12733	-1.69	0.28	
12707	-0.30	-0.02	12706	-1.64	-0.12	
13911	-0.58	-0.01	11546	-1.50	-0.21	
12954	-1.97	-0.01	12830	-1.17	-0.10	
12971	0.35	0.00	16652	-1.17	0.06	
16653	0.97	0.00	12957	-1.12	0.12	
13912	-0.24	0.01	12988	-1.07	-0.10	
16604	-0.07	0.02	13789	-0.72	-0.28	
13906	-2.24	0.02	13911	-0.58	-0.01	
16686	-0.47	0.03	16686	-0.47	0.03	
12787	-0.11	0.03	16603	-0.35	-0.03	
16684	2.02	0.04	12707	-0.30	-0.02	
14884	-2.64	0.05	13912	-0.24	0.01	
13910	-2.33	0.05	12705	-0.17	-0.04	
16652	-1.17	0.06	12787	-0.11	0.03	
17769	-4.25	0.06	16604	-0.07	0.02	
14883	0.51	0.08	12991	-0.07	0.22	
13904	-3.24	0.09	12785	0.00	-0.04	
11544	-3.56	0.09	12953	0.13	0.22	
12957	-1.12	0.12	12734	0.26	-0.15	
13905	-3.26	0.14	12990	0.29	0.24	
12956	-3.09	0.15	12971	0.35	0.00	
12955	1.23	0.16	14883	0.51	0.08	
17770	-2.72	0.18	16685	0.53	-0.03	
14879	-4.12	0.21	16653	0.97	0.00	
12991	-0.07	0.22	16654	1.15	-0.04	
12735	1.28	0.22	12955	1.23	0.16	
12953	0.13	0.22	12735	1.28	0.22	
12990	0.29	0.24	12786	1.56	-0.11	
12733	-1.69	0.28	16684	2.02	0.04	

2.7.1.3 Grades 4–5

Table 2.7.1.3

Equating Summary: List 4-5 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	54	0.57 (1.58)		54	0.42 (1.66)	
	Easiest	Hardest		Easiest	Hardest	
	-2.65	4.13		-2.65	4.13	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	42	0.49 (1.71)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	39	0.55 (1.73)				
	Percentage Anchors	Average Displacement				
72%	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
	13027	3.33	-0.25	12919	-2.65	0.07
	13028	3.20	-0.24	12793	-2.36	-0.06
	14939	2.28	-0.18	16618	-2.08	0.20
	13024	-1.78	-0.16	12792	-1.97	0.25
	14945	0.10	-0.15	13024	-1.78	-0.16
	17789	0.08	-0.12	17710	-1.51	0.01
	13025	-0.05	-0.11	16613	-0.78	0.12
	17714	-0.52	-0.11	13026	-0.76	0.19
	12794	-0.22	-0.09	16708	-0.72	-0.01
	16709	1.19	-0.07	16710	-0.52	-0.03
	12793	-2.36	-0.06	17714	-0.52	-0.11
	14946	0.95	-0.06	12918	-0.47	0.05
	16710	-0.52	-0.03	12917	-0.39	0.17
	16616	0.29	-0.03	12794	-0.22	-0.09
	14214	4.13	-0.02	13025	-0.05	-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
	14940	0.48	-0.02	14947	-0.04	0.15
16619	2.15	-0.01	17789	0.08	-0.12	
12942	1.74	-0.01	14945	0.10	-0.15	
16708	-0.72	-0.01	16615	0.10	0.05	
13029	1.83	0.00	16616	0.29	-0.03	
14212	2.49	0.00	14941	0.32	0.15	
17710	-1.51	0.01	14940	0.48	-0.02	
12918	-0.47	0.05	16712	0.50	0.25	
16615	0.10	0.05	14946	0.95	-0.06	
12946	1.46	0.06	14213	1.13	0.10	
12919	-2.65	0.07	16709	1.19	-0.07	
16620	2.50	0.08	12946	1.46	0.06	
16714	2.68	0.09	16713	1.64	0.16	
14213	1.13	0.10	12942	1.74	-0.01	
16613	-0.78	0.12	13029	1.83	0.00	
14941	0.32	0.15	16619	2.15	-0.01	
14947	-0.04	0.15	14939	2.28	-0.18	
16713	1.64	0.16	14212	2.49	0.00	
12917	-0.39	0.17	16620	2.50	0.08	
13026	-0.76	0.19	16714	2.68	0.09	
16618	-2.08	0.20	13028	3.20	-0.24	
12943	3.60	0.24	13027	3.33	-0.25	
16712	0.50	0.25	12943	3.60	0.24	
12792	-1.97	0.25	14214	4.13	-0.02	

2.7.1.4 Grades 6–8

Table 2.7.1.4

Equating Summary: List 6-8 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	54	1.33 (1.27)		54	1.26 (1.23)	
	Easiest	Hardest		Easiest	Hardest	
	-1.14	3.55		-0.83	3.55	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	42	1.43 (1.21)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	37	1.49 (1.15)				
	Percentage Anchors	Average Displacement				
69%	-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
	16563	0.06	-0.29	13040	-0.83	0.07
	14863	1.43	-0.24	13039	-0.52	0.07
	14856	3.31	-0.23	13041	-0.29	-0.02
	17686	0.88	-0.22	17678	-0.20	0.02
	14851	1.65	-0.16	16563	0.06	-0.29
	14855	2.29	-0.15	16664	0.10	0.07
	14917	1.05	-0.13	16562	0.24	0.06
	16568	2.04	-0.10	14850	0.34	0.02
	14857	3.55	-0.10	17686	0.88	-0.22
	14915	1.08	-0.10	17694	0.89	-0.06
	17694	0.89	-0.06	14917	1.05	-0.13
	16566	1.86	-0.05	17725	1.07	0.16
	16665	1.13	-0.03	14915	1.08	-0.10
	14923	3.37	-0.03	16665	1.13	-0.03
	13041	-0.29	-0.02	17687	1.21	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
	14922	2.74	-0.02	16666	1.39	0.03
14921	2.48	-0.01	14863	1.43	-0.24	
14916	1.73	-0.01	14859	1.51	0.00	
14858	1.54	-0.01	17726	1.54	0.16	
14852	1.81	-0.01	14858	1.54	-0.01	
16674	3.45	0.00	16564	1.63	0.03	
14859	1.51	0.00	14851	1.65	-0.16	
17678	-0.20	0.02	14916	1.73	-0.01	
14850	0.34	0.02	17688	1.79	0.04	
16666	1.39	0.03	14852	1.81	-0.01	
16564	1.63	0.03	16566	1.86	-0.05	
17688	1.79	0.04	16568	2.04	-0.10	
16562	0.24	0.06	16673	2.13	0.13	
13039	-0.52	0.07	14855	2.29	-0.15	
16664	0.10	0.07	14921	2.48	-0.01	
13040	-0.83	0.07	14922	2.74	-0.02	
17687	1.21	0.08	16672	2.78	0.10	
16672	2.78	0.10	16567	3.07	0.26	
16673	2.13	0.13	14856	3.31	-0.23	
17726	1.54	0.16	14923	3.37	-0.03	
17725	1.07	0.16	16674	3.45	0.00	
16567	3.07	0.26	14857	3.55	-0.10	

2.7.1.5 Grades 9-12

Table 2.7.1.5

Equating Summary: List 9-12 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	54	1.52 (1.24)		54	1.53 (1.16)	
	Easiest	Hardest		Easiest	Hardest	
	-0.80	4.08		-0.80	3.88	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	48	1.63 (1.24)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	36	1.59 (1.21)				
	Percentage Anchors	Average Displacement				
67%	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
	16656	2.18	-0.20	12887	-0.80	0.28
	17761	-0.38	-0.18	12716	-0.45	-0.06
	16587	2.22	-0.16	17761	-0.38	-0.18
	16658	2.48	-0.14	12714	0.00	0.09
	13867	3.88	-0.13	12893	0.07	0.06
	16586	1.08	-0.13	12889	0.12	0.10
	17721	2.25	-0.12	14954	0.53	0.22
	14885	2.16	-0.10	17741	0.54	0.13
	16657	1.04	-0.10	12890	0.60	0.00
	13038	3.60	-0.09	16588	0.92	-0.01
	14886	1.96	-0.09	17762	0.94	0.06
	12895	2.83	-0.08	12715	1.01	0.01
	12716	-0.45	-0.06	16657	1.04	-0.10
	14956	1.86	-0.04	13036	1.07	0.20
	13037	2.66	-0.03	16586	1.08	-0.13

Displacement of Anchor Items	Anchor Items by Displacement			Anchor Items by Item Difficulty		
	Item ID	Item Difficulty	Displacement	Item ID	Item Difficulty	Displacement
	16588	0.92	-0.01	17742	1.34	0.13
12890	0.60	0.00	16591	1.54	0.13	
12715	1.01	0.01	17763	1.62	0.09	
16590	2.17	0.03	17743	1.77	0.05	
13866	3.49	0.04	14956	1.86	-0.04	
14955	1.98	0.04	14886	1.96	-0.09	
13865	2.68	0.05	14955	1.98	0.04	
16592	3.61	0.05	14885	2.16	-0.10	
17743	1.77	0.05	16590	2.17	0.03	
17762	0.94	0.06	16656	2.18	-0.20	
12893	0.07	0.06	16587	2.22	-0.16	
17763	1.62	0.09	17721	2.25	-0.12	
12714	0.00	0.09	16658	2.48	-0.14	
12889	0.12	0.10	13037	2.66	-0.03	
12894	2.82	0.11	13865	2.68	0.05	
17742	1.34	0.13	12894	2.82	0.11	
17741	0.54	0.13	12895	2.83	-0.08	
16591	1.54	0.13	13866	3.49	0.04	
13036	1.07	0.20	13038	3.60	-0.09	
14954	0.53	0.22	16592	3.61	0.05	
12887	-0.80	0.28	13867	3.88	-0.13	

2.7.2 Reading

2.7.2.1 Grade 1

Table 2.7.2.1

Equating Summary: Read 1 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	72	-1.00 (1.12)		72	-1.06 (1.01)	
	Easiest	Hardest		Easiest	Hardest	
	-4.07	1.01		-4.32	1.01	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	57	-0.84 (1.09)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	42	-0.75 (0.96)				
	Percentage Anchors	Average Displacement				
58%	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
	17135	-0.85	-0.20	13274	-4.07	-0.06
	16055	-0.18	-0.18	15681	-2.77	0.17
	16054	0.27	-0.18	13193	-2.11	-0.01
	16041	-1.07	-0.13	13194	-2.06	0.05
	16039	-1.61	-0.12	16039	-1.61	-0.12
	13245	-0.81	-0.12	17035	-1.58	0.06
	13195	-1.52	-0.11	13238	-1.58	-0.05
	17956	-0.74	-0.09	13195	-1.52	-0.11
	17131	-0.34	-0.08	17033	-1.14	0.03
	13274	-4.07	-0.06	13276	-1.13	0.15
	17029	0.83	-0.06	13275	-1.10	0.20
	14621	0.20	-0.05	16041	-1.07	-0.13
	16040	-1.02	-0.05	13217	-1.03	0.07
	13238	-1.58	-0.05	16040	-1.02	-0.05
13240	-0.95	-0.04	13240	-0.95	-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
	16053	1.01	-0.03	17138	-0.94	0.00
13246	-0.79	-0.02	17034	-0.94	0.04	
17133	0.84	-0.02	17135	-0.85	-0.20	
14619	-0.05	-0.02	13245	-0.81	-0.12	
17031	0.51	-0.02	13244	-0.80	0.03	
13193	-2.11	-0.01	13246	-0.79	-0.02	
17138	-0.94	0.00	17956	-0.74	-0.09	
17955	-0.54	0.01	17986	-0.74	0.10	
13244	-0.80	0.03	17139	-0.67	0.09	
17033	-1.14	0.03	13239	-0.60	0.12	
17034	-0.94	0.04	17955	-0.54	0.01	
13194	-2.06	0.05	13218	-0.45	0.13	
17035	-1.58	0.06	17132	-0.43	0.12	
13217	-1.03	0.07	17131	-0.34	-0.08	
17139	-0.67	0.09	17958	-0.34	0.21	
17986	-0.74	0.10	17030	-0.30	0.24	
17132	-0.43	0.12	16055	-0.18	-0.18	
13239	-0.60	0.12	14619	-0.05	-0.02	
13218	-0.45	0.13	14620	0.03	0.22	
13276	-1.13	0.15	13219	0.05	0.16	
13219	0.05	0.16	17987	0.09	0.25	
15681	-2.77	0.17	14621	0.20	-0.05	
13275	-1.10	0.20	16054	0.27	-0.18	
17958	-0.34	0.21	17031	0.51	-0.02	
14620	0.03	0.22	17029	0.83	-0.06	
17030	-0.30	0.24	17133	0.84	-0.02	
17987	0.09	0.25	16053	1.01	-0.03	

2.7.2.2 Grades 2–3

Table 2.7.2.2

Equating Summary: Read 2-3 S502 Online

Comparison of Forms	Form 502		Form 501			
	No. of Items	Average Difficulty (Std. Dev.)	No. of Items	Average Difficulty (Std. Dev.)		
	72	-0.07 (0.94)	72	-0.10 (0.99)		
	Easiest	Hardest	Easiest	Hardest		
-2.97	2.46	-2.97	2.06			
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	60	-0.08 (0.98)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	45	0.09 (0.88)				
	Percentage Anchors	Average Displacement				
63%	-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
	17043	0.08	-0.28	13353	-2.97	0.03
	13346	0.60	-0.26	15626	-1.69	0.11
	17041	-0.20	-0.25	13331	-1.45	0.14
	13375	1.34	-0.24	13393	-1.25	-0.03
	15716	0.36	-0.23	15700	-0.83	0.17
	13344	0.33	-0.21	15628	-0.79	0.13
	13338	0.80	-0.18	13355	-0.57	-0.14
	17154	-0.04	-0.17	13279	-0.51	-0.03
	13355	-0.57	-0.14	17153	-0.51	-0.08
	13340	-0.25	-0.12	17950	-0.45	0.07
	16092	1.27	-0.10	17878	-0.44	0.14
	13345	1.24	-0.10	14589	-0.34	0.12
	16095	0.70	-0.10	13333	-0.34	-0.02
	13339	0.38	-0.09	13374	-0.30	0.14
17049	1.22	-0.08	13340	-0.25	-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
	17153	-0.51	-0.08	14590	-0.21	0.02
13376	0.66	-0.07	17041	-0.20	-0.25	
17042	0.08	-0.05	15715	-0.16	0.01	
17155	0.17	-0.05	13932	-0.08	0.04	
17050	1.25	-0.04	17154	-0.04	-0.17	
16094	0.90	-0.04	17043	0.08	-0.28	
13279	-0.51	-0.03	17042	0.08	-0.05	
13393	-1.25	-0.03	17155	0.17	-0.05	
13333	-0.34	-0.02	17952	0.28	0.15	
15715	-0.16	0.01	17051	0.32	0.01	
15629	0.42	0.01	13344	0.33	-0.21	
17051	0.32	0.01	15716	0.36	-0.23	
14590	-0.21	0.02	13339	0.38	-0.09	
13353	-2.97	0.03	15629	0.42	0.01	
13932	-0.08	0.04	13937	0.42	0.12	
17950	-0.45	0.07	13936	0.45	0.16	
14591	0.81	0.11	13346	0.60	-0.26	
15626	-1.69	0.11	13376	0.66	-0.07	
13937	0.42	0.12	16095	0.70	-0.10	
14589	-0.34	0.12	13338	0.80	-0.18	
15628	-0.79	0.13	14591	0.81	0.11	
13374	-0.30	0.14	17951	0.87	0.18	
17878	-0.44	0.14	16094	0.90	-0.04	
13331	-1.45	0.14	17924	1.13	0.23	
17952	0.28	0.15	17049	1.22	-0.08	
13936	0.45	0.16	13345	1.24	-0.10	
15700	-0.83	0.17	17050	1.25	-0.04	
17951	0.87	0.18	16092	1.27	-0.10	
17924	1.13	0.23	17934	1.31	0.28	
17934	1.31	0.28	13375	1.34	-0.24	

2.7.2.3 Grades 4–5

Table 2.7.2.3

Equating Summary: Read 4-5 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	72	1.00 (1.11)		72	0.90 (1.27)	
	Easiest	Hardest		Easiest	Hardest	
	-1.07	3.49		-3.19	3.49	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	57	1.05 (1.11)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	49	1.19 (1.08)				
	Percentage Anchors	Average Displacement				
68%	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
	16019	1.59	-0.28	14715	-0.75	-0.09
	13504	2.57	-0.25	13407	-0.72	-0.21
	17110	1.17	-0.23	18184	-0.52	-0.04
	13407	-0.72	-0.21	14714	-0.37	-0.03
	18198	0.71	-0.20	14626	-0.26	0.18
	13505	1.96	-0.18	13409	-0.17	0.01
	16011	0.12	-0.18	16009	-0.10	-0.04
	15706	0.21	-0.17	17109	-0.03	0.00
	18186	0.66	-0.16	13408	0.06	0.08
	15708	1.38	-0.16	16017	0.09	-0.11
	16017	0.09	-0.11	16011	0.12	-0.18
	14715	-0.75	-0.09	14716	0.20	0.04
	13528	1.64	-0.09	15706	0.21	-0.17
	15707	1.23	-0.08	14625	0.26	-0.06
	14625	0.26	-0.06	18193	0.52	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
	16018	1.28	-0.06	18186	0.66	-0.16
18185	1.15	-0.04	18198	0.71	-0.20	
16009	-0.10	-0.04	18197	0.80	0.02	
18184	-0.52	-0.04	18125	0.98	0.14	
16010	1.34	-0.04	13529	0.98	0.13	
13527	1.08	-0.04	18128	0.99	0.26	
17113	2.64	-0.04	13527	1.08	-0.04	
14714	-0.37	-0.03	18185	1.15	-0.04	
13928	3.31	-0.03	17110	1.17	-0.23	
17109	-0.03	0.00	15707	1.23	-0.08	
18200	1.80	0.00	16018	1.28	-0.06	
17111	1.47	0.01	16010	1.34	-0.04	
13409	-0.17	0.01	15708	1.38	-0.16	
18197	0.80	0.02	18123	1.39	0.09	
13482	2.46	0.02	17111	1.47	0.01	
14716	0.20	0.04	13926	1.50	0.28	
13484	2.41	0.04	14627	1.59	0.16	
13927	2.81	0.05	16019	1.59	-0.28	
18201	2.63	0.05	13528	1.64	-0.09	
18194	3.49	0.07	18200	1.80	0.00	
18193	0.52	0.07	13505	1.96	-0.18	
18192	2.16	0.07	13483	2.13	0.26	
13408	0.06	0.08	17112	2.14	0.29	
18123	1.39	0.09	13503	2.15	0.12	
13503	2.15	0.12	18192	2.16	0.07	
13529	0.98	0.13	18202	2.37	0.15	
18125	0.98	0.14	13484	2.41	0.04	
18202	2.37	0.15	13482	2.46	0.02	
14627	1.59	0.16	13504	2.57	-0.25	
14626	-0.26	0.18	18201	2.63	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
	18128	0.99	0.26	17113	2.64	-0.04
13483	2.13	0.26	13927	2.81	0.05	
13926	1.50	0.28	13928	3.31	-0.03	
17112	2.14	0.29	18194	3.49	0.07	

2.7.2.4 Grades 6–8

Table 2.7.2.4

Equating Summary: Read 6-8 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	72	1.38 (1.41)		72	1.30 (1.42)	
	Easiest	Hardest		Easiest	Hardest	
	-1.69	4.05		-1.69	4.05	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	57	1.42 (1.49)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	51	1.38 (1.54)				
	Percentage Anchors	Average Displacement				
71%	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
	13963	1.94	-0.21	18062	-1.69	0.04
	17015	2.41	-0.20	13575	-1.58	0.16
	13657	1.62	-0.20	13565	-1.25	0.28
	18055	1.14	-0.20	14641	-1.23	0.14
	18091	0.37	-0.20	15713	-1.16	-0.03
	13962	1.39	-0.18	13564	-0.98	0.26
	13629	0.78	-0.16	14640	-0.55	-0.05
	17127	2.11	-0.12	13563	-0.52	0.16
	13658	3.59	-0.09	13576	-0.48	0.25
	17124	3.02	-0.07	15712	-0.38	0.20
	17014	2.20	-0.06	13577	-0.09	0.22
	17027	3.21	-0.06	15714	0.19	0.11
	14640	-0.55	-0.05	18091	0.37	-0.20
	18064	0.57	-0.05	18063	0.44	-0.02
	14618	2.75	-0.05	18092	0.48	-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
	13650	1.09	-0.05	14642	0.52	0.19
17129	2.75	-0.05	18064	0.57	-0.05	
13964	2.02	-0.04	13629	0.78	-0.16	
17128	2.21	-0.03	13631	0.84	0.00	
14616	2.52	-0.03	13650	1.09	-0.05	
15713	-1.16	-0.03	18055	1.14	-0.20	
18063	0.44	-0.02	13661	1.22	0.08	
17025	3.83	-0.02	13659	1.26	0.10	
18092	0.48	-0.01	13962	1.39	-0.18	
17123	3.94	-0.01	13657	1.62	-0.20	
13651	2.53	0.00	13660	1.89	0.11	
13631	0.84	0.00	13963	1.94	-0.21	
18062	-1.69	0.04	13614	1.95	0.07	
13615	2.21	0.04	18076	1.99	0.15	
13616	2.67	0.04	13964	2.02	-0.04	
13652	2.41	0.04	17026	2.04	0.15	
13656	2.35	0.05	18074	2.06	0.30	
14617	3.66	0.06	17127	2.11	-0.12	
17125	4.05	0.07	17014	2.20	-0.06	
13614	1.95	0.07	13615	2.21	0.04	
13661	1.22	0.08	17128	2.21	-0.03	
13659	1.26	0.10	13656	2.35	0.05	
15714	0.19	0.11	17015	2.41	-0.20	
13660	1.89	0.11	13652	2.41	0.04	
14641	-1.23	0.14	14616	2.52	-0.03	
17026	2.04	0.15	13651	2.53	0.00	
18076	1.99	0.15	13616	2.67	0.04	
13563	-0.52	0.16	14618	2.75	-0.05	
13575	-1.58	0.16	17129	2.75	-0.05	
14642	0.52	0.19	17124	3.02	-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
	15712	-0.38	0.20	17027	3.21	-0.06
13577	-0.09	0.22	13658	3.59	-0.09	
13576	-0.48	0.25	14617	3.66	0.06	
13564	-0.98	0.26	17025	3.83	-0.02	
13565	-1.25	0.28	17123	3.94	-0.01	
18074	2.06	0.30	17125	4.05	0.07	

2.7.2.5 Grades 9-12

Table 2.7.2.5

Equating Summary: Read 9-12 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Items	Average Difficulty (Std. Dev.)		No. of Items	Average Difficulty (Std. Dev.)	
	72	2.23 (1.20)		72	2.08 (1.26)	
	Easiest	Hardest		Easiest	Hardest	
	-1.20	4.52		-1.77	4.52	
Anchoring Items	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	51	2.41 (1.12)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	48	2.45 (1.13)				
	Percentage Anchors	Average Displacement				
67%	-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
	14976	2.99	-0.24	17998	0.45	-0.05
	13950	1.41	-0.21	16062	0.53	-0.13
	16064	1.45	-0.19	17996	0.59	-0.02
	14975	1.87	-0.17	16058	0.62	-0.13
	17938	0.63	-0.15	17938	0.63	-0.15
	16062	0.53	-0.13	16059	0.67	0.01
	16058	0.62	-0.13	17939	0.75	0.12
	17071	2.49	-0.12	16063	0.98	-0.04
	16072	3.26	-0.11	18023	1.11	0.01
	13786	4.30	-0.10	13950	1.41	-0.21
	14977	4.01	-0.10	16064	1.45	-0.19
	17940	2.07	-0.10	17933	1.84	0.19
	13733	2.65	-0.10	14975	1.87	-0.17
	16070	2.65	-0.08	18025	1.88	0.03
	14635	3.35	-0.07	17079	1.91	-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
	17091	3.14	-0.07	13952	1.92	0.01
17998	0.45	-0.05	17940	2.07	-0.10	
13731	3.95	-0.05	14634	2.09	0.13	
16063	0.98	-0.04	17936	2.16	0.06	
13951	2.69	-0.03	18030	2.25	0.10	
17079	1.91	-0.03	16071	2.36	0.02	
17075	3.98	-0.02	17076	2.40	0.16	
17996	0.59	-0.02	17073	2.42	0.01	
17077	4.52	-0.01	17071	2.49	-0.12	
17080	3.06	-0.01	17935	2.50	0.00	
17072	2.93	0.00	13785	2.54	0.18	
14636	3.11	0.00	13733	2.65	-0.10	
17935	2.50	0.00	16070	2.65	-0.08	
16059	0.67	0.01	13951	2.69	-0.03	
17073	2.42	0.01	13732	2.92	0.16	
18023	1.11	0.01	17072	2.93	0.00	
13952	1.92	0.01	14976	2.99	-0.24	
16071	2.36	0.02	17080	3.06	-0.01	
18025	1.88	0.03	14636	3.11	0.00	
17081	3.28	0.03	17091	3.14	-0.07	
17093	4.05	0.04	16072	3.26	-0.11	
17092	4.13	0.04	17081	3.28	0.03	
13787	3.69	0.04	14635	3.35	-0.07	
17936	2.16	0.06	18032	3.45	0.16	
18030	2.25	0.10	18031	3.57	0.14	
17939	0.75	0.12	13787	3.69	0.04	
14634	2.09	0.13	13731	3.95	-0.05	
18031	3.57	0.14	17075	3.98	-0.02	
18032	3.45	0.16	14977	4.01	-0.10	
13732	2.92	0.16	17093	4.05	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
	17076	2.40	0.16	17092	4.13	0.04
	13785	2.54	0.18	13786	4.30	-0.10
17933	1.84	0.19	17077	4.52	-0.01	

2.7.3 Writing

2.7.3.1 Grade 1

Table 2.7.3.1.1

Equating Summary: Writ 1 A S502 Online

Comparison of Forms	Form 502			Form 501					
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)				
	2	-0.22 (0.07)		2	-0.18 (0.13)				
	Easiest	Hardest		Easiest	Hardest				
	-0.27	-0.18		-0.27	-0.08				
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)							
	1	-0.27 (N/A)							
	No. of Anchors Used	Average Difficulty (Std. Dev.)							
	1	-0.27 (N/A)							
	Percentage Anchors	Average Displacement							
50%	0.01								
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	Difficulty	Displacement	Task ID	Difficulty	Displacement			
	14248	-0.27	0.01	14248	-0.27	0.01			

Table 2.7.3.1.2

Equating Summary: Writ 1 B/C S502 Online

Comparison of Forms	Form 502		Form 501			
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)		
	2	0.28 (0.17)	2	0.32 (0.12)		
	Easiest 0.16	Hardest 0.40	Easiest 0.24	Hardest 0.40		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	1	0.40 (N/A)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	1	0.40 (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
	18231	0.40	0.00	18231	0.40	0.00

2.7.3.2 Grades 2–3

Table 2.7.3.2.1

Equating Summary: Writ 2-3 A S502 Online

Comparison of Forms	Form 502		Form 501			
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)		
	2	-0.10 (0.24)	2	-0.06 (0.30)		
	Easiest	Hardest	Easiest	Hardest		
	-0.27	0.07	-0.27	0.16		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	1	-0.27 (N/A)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	1	-0.27 (N/A)				
	Percentage Anchors	Average Displacement				
50%	-0.02					
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
	18232	-0.27	-0.02	18232	-0.27	-0.02

Table 2.7.3.2.2

Equating Summary: Writ 2-3 B/C S502 Online

Comparison of Forms	Form 502		Form 501					
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)				
	2	0.05 (0.26)	2	0.43 (0.80)				
	Easiest	Hardest	Easiest	Hardest				
	-0.14	0.23	-0.14	1.00				
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)						
	1	-0.14 (N/A)						
	No. of Anchors Used	Average Difficulty (Std. Dev.)						
	1	-0.14 (N/A)						
	Percentage Anchors	Average Displacement						
50%	-0.04							
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		
	17304	-0.14	-0.04	17304	-0.14	-0.04		

2.7.3.3 Grades 4–5

Table 2.7.3.3.1

Equating Summary: Writ 4-5 A S502 Online

Comparison of Forms	Form 502		Form 501			
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)		
	2	1.10 (0.15)	2	1.19 (0.28)		
	Easiest 0.99	Hardest 1.21	Easiest 0.99	Hardest 1.39		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	1	0.99 (N/A)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	1	0.99 (N/A)				
	Percentage Anchors	Average Displacement				
50%	0.01					
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
	6	0.99	0.01	17668_18236	0.99	0.01

Table 2.7.3.3.2

Equating Summary: Writ 4-5 B/C S502 Online

Comparison of Forms	Form 502		Form 501			
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)		
	2	2.55 (0.09)	2	2.48 (0.18)		
	Easiest 2.49	Hardest 2.61	Easiest 2.35	Hardest 2.61		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	1	2.61 (N/A)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	1	2.61 (N/A)				
	Percentage Anchors	Average Displacement				
50%	-0.03					
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
	17328_18238	2.61	-0.03	17328_18238	2.61	-0.03

2.7.3.4 Grades 6–8

Table 2.7.3.4.1

Equating Summary: Writ 6-8 A S502 Online

Comparison of Forms	Form 502		Form 501					
	No. of Tasks	Average Difficulty (Std. Dev.)	No. of Tasks	Average Difficulty (Std. Dev.)				
	2	0.94 (0.06)	2	1.30 (0.45)				
	Easiest 0.90	Hardest 0.99	Easiest 0.99	Hardest 1.62				
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)						
	1	0.99 (N/A)						
	No. of Anchors Used	Average Difficulty (Std. Dev.)						
	1	0.99 (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		
	17347_18243	0.99	0.00	17347_18243	0.99	0.00		

Table 2.7.3.4.2

Equating Summary: Writ 6-8 B/C S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)	
	2	1.52 (0.00)		2	1.51 (0.02)	
	Easiest	Hardest		Easiest	Hardest	
1.52	1.52		1.49	1.52		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	1	1.52 (N/A)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	1	1.52 (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
	16388_17262	1.52	0.00	16388_17262	1.52	0.00

2.7.3.5 Grades 9-12

Table 2.7.3.5.1

Equating Summary: Writ 9-12 A S502 Online

Comparison of Forms	Form 502			Form 501					
	No. of Tasks		Average Difficulty (Std. Dev.)	No. of Tasks		Average Difficulty (Std. Dev.)			
	2		2.05 (0.37)	2		1.92 (0.18)			
	Easiest		Hardest	Easiest		Hardest			
1.79		2.31	1.79		2.05				
Anchoring Tasks	No. of Possible Anchors		Average Difficulty (Std. Dev.)						
	1		1.79 (N/A)						
	No. of Anchors Used		Average Difficulty (Std. Dev.)						
	1		1.79 (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			
	17333_18250	1.79	0.00	0	1.79	0.00			

Table 2.7.3.5.2

Equating Summary: Writ 9-12 B/C S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)	
	2	2.02 (0.42)		2	2.07 (0.35)	
	Easiest	Hardest		Easiest	Hardest	
	1.72	2.32		1.83	2.32	
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%	0.04					
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	0.04	17319_18252	2.32	0.04

2.7.4 Speaking

2.7.4.1 Grade 1

Table 2.7.4.1

Equating Summary: Spek 1 S502 Online

Comparison of Forms	Form 502			Form 501				
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)			
	9	-1.64 (2.06)		9	-1.48 (2.30)			
	Easiest	Hardest		Easiest	Hardest			
	-4.45	0.31		-4.70	0.82			
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)						
	3	-1.56 (2.47)						
	No. of Anchors Used	Average Difficulty (Std. Dev.)						
	3	-1.56 (2.47)						
	Percentage Anchors	Average Displacement						
33%	0.13							
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		
	18749	-4.45	0.00	18749	-4.45	0.00		
	18754	-0.72	0.00	17460	-4.42	0.05		
	18723	-4.22	0.00	18723	-4.22	0.00		
	18727	-0.35	0.00	18754	-0.72	0.00		
	18758	0.31	0.00	18733	-0.65	0.00		
	18733	-0.65	0.00	18727	-0.35	0.00		
	17460	-4.42	0.05	17465	-0.27	0.09		
	17465	-0.27	0.09	17470	0.00	0.24		
	17470	0.00	0.24	18758	0.31	0.00		

2.7.4.2 Grades 2–3

Table 2.7.4.2

Equating Summary: Spek 2-3 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)	
	9	-1.61 (2.27)		9	-1.61 (2.40)	
	Easiest	Hardest		Easiest	Hardest	
-4.81	0.47		-4.81	0.95		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	3	-1.55 (2.85)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	3	-1.55 (2.85)				
	Percentage Anchors	Average Displacement				
33%	0.01					
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
	15068	-0.29	-0.17	15055	-4.81	0.18
	18761	-4.46	0.00	18773	-4.51	0.00
	18766	-0.25	0.00	18761	-4.46	0.00
	18773	-4.51	0.00	18778	-0.75	0.00
	18778	-0.75	0.00	18785	-0.35	0.00
	18770	0.47	0.00	15068	-0.29	-0.17
	18785	-0.35	0.00	18766	-0.25	0.00
	15081	0.46	0.03	15081	0.46	0.03
15055	-4.81	0.18	18770	0.47	0.00	

2.7.4.3 Grades 4–5

Table 2.7.4.3

Equating Summary: Spek 4-5 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)	
	9	-0.33 (2.36)		9	-0.35 (2.81)	
	Easiest	Hardest		Easiest	Hardest	
-3.58	1.98		-4.34	2.15		
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	3	-0.52 (2.71)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	2	-0.52 (2.71)				
	Percentage Anchors	Average Displacement				
22%	0.06					
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
	17530	0.45	-0.06	17524	-3.58	0.00
	18830	-3.37	0.00	18830	-3.37	0.00
	18835	0.91	0.00	18816	-3.34	0.00
	18816	-3.34	0.00	17530	0.45	-0.06
	18821	0.90	0.00	18821	0.90	0.00
	17524	-3.58	0.00	18835	0.91	0.00
	18839	1.47	0.00	18839	1.47	0.00
	18827	1.98	0.00	17535	1.57	0.23
17535	1.57	0.23	18827	1.98	0.00	

2.7.4.4 Grades 6–8

Table 2.7.4.4

Equating Summary: Spek 6-8 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)	
	9	0.18 (2.46)		9	0.19 (2.64)	
	Easiest	Hardest		Easiest	Hardest	
	-3.17	2.18		-3.52	2.53	
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	3	0.19 (2.92)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	3	0.19 (2.92)				
	Percentage Anchors	Average Displacement				
33%	0.05					
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
	17609	1.69	-0.04	17604	-3.17	0.03
	18636	-3.04	0.00	18636	-3.04	0.00
	18641	1.25	0.00	18662	-3.00	0.00
	18662	-3.00	0.00	18641	1.25	0.00
	18667	1.56	0.00	18667	1.56	0.00
	18646	2.18	0.00	17609	1.69	-0.04
	18672	2.12	0.00	17614	2.06	0.17
	17604	-3.17	0.03	18672	2.12	0.00
17614	2.06	0.17	18646	2.18	0.00	

2.7.4.5 Grades 9-12

Table 2.7.4.5

Equating Summary: Spek 9-12 S502 Online

Comparison of Forms	Form 502			Form 501		
	No. of Tasks	Average Difficulty (Std. Dev.)		No. of Tasks	Average Difficulty (Std. Dev.)	
	9	0.58 (2.49)		9	0.41 (2.50)	
	Easiest	Hardest		Easiest	Hardest	
	-2.76	3.02		-3.13	2.71	
Anchoring Tasks	No. of Possible Anchors	Average Difficulty (Std. Dev.)				
	3	0.49 (2.83)				
	No. of Anchors Used	Average Difficulty (Std. Dev.)				
	3	0.49 (2.83)				
	Percentage Anchors	Average Displacement				
33%	-0.12					
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
	17576	2.35	-0.26	17564	-2.76	-0.09
	17564	-2.76	-0.09	17592	-2.76	0.00
	17570	1.88	-0.02	18688	-2.61	0.00
	18688	-2.61	0.00	18693	1.87	0.00
	18693	1.87	0.00	17570	1.88	-0.02
	17592	-2.76	0.00	17598	2.03	0.00
	17598	2.03	0.00	18699	2.16	0.00
	18699	2.16	0.00	17576	2.35	-0.26
	18233	3.02	0.00	18233	3.02	0.00

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.

In addition to the TCC by tier, we also plotted on the same graph the TCCs across tiers for the grade-level clusters. Since the ranges of raw score points that a student can achieve differ depending upon the tier of the test form that the student takes, it is not appropriate to compare the expected raw score points for the same student ability measure for different tiers. It is, however, informative to compare where the slopes are the steepest, which corresponds to the ability range that provides more measurement information (i.e., better targeting of the task to the abilities of students in that range, and lower standard errors for the student ability measures in that range). For example, as shown in Figure 2.8.3.1.3, the across-tiers TCC for Writing Grade 1 indicated that the Writing Tier A form provided more measurement of writing ability for those students who had ability measures at -2.5 logits and 4.0 logits, whereas the Writing Grade 1 Tier B/C form provided more measurement of writing ability for those students who had ability measures at -2.0 logits and 4.5 logits.

In addition, it is informative to compare the area under the curve of the TCC for each tier form. For example, as shown in Figure 2.8.3.1.3, the Grade 1 Tier A curve for Writing covers a larger area of the lower ability range than the Grade 1 Tier B/C curve, especially at the very low end of the student ability continuum. Consistent with the purposes of the test design, there is also considerable overlap between the ranges of writing ability that the two forms cover.

2.8.1 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

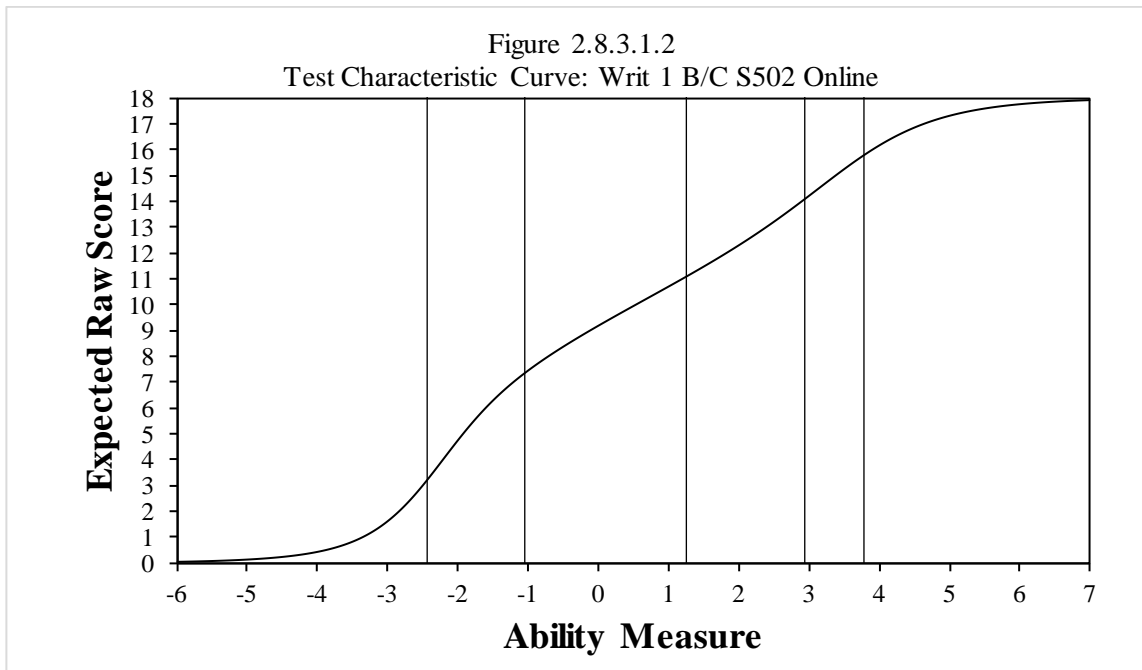
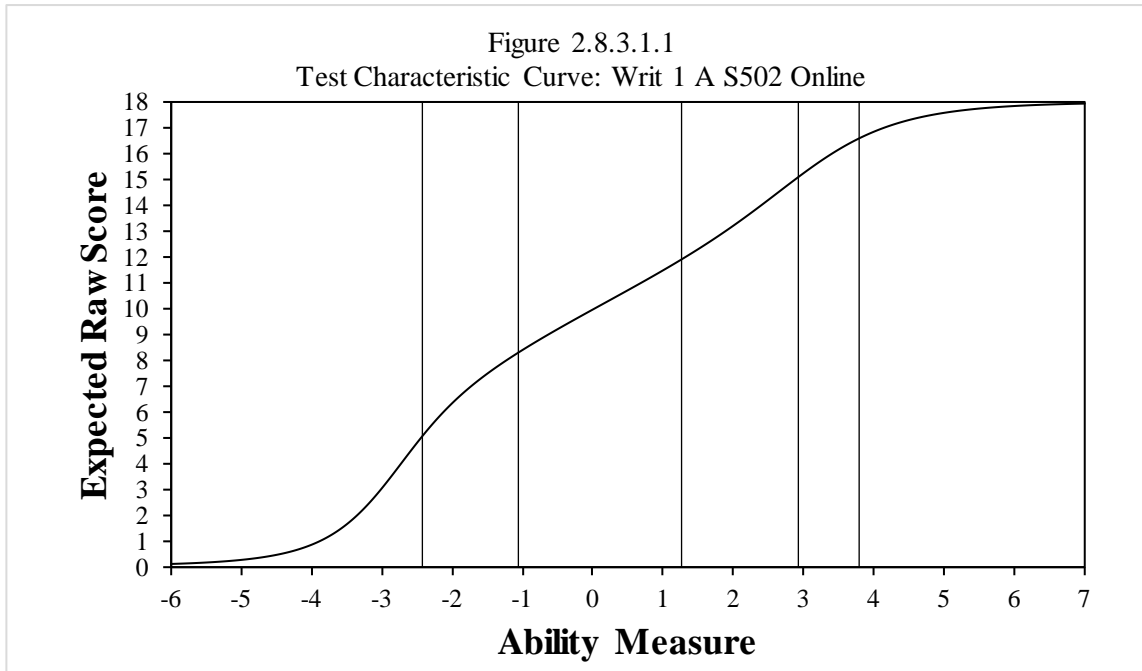
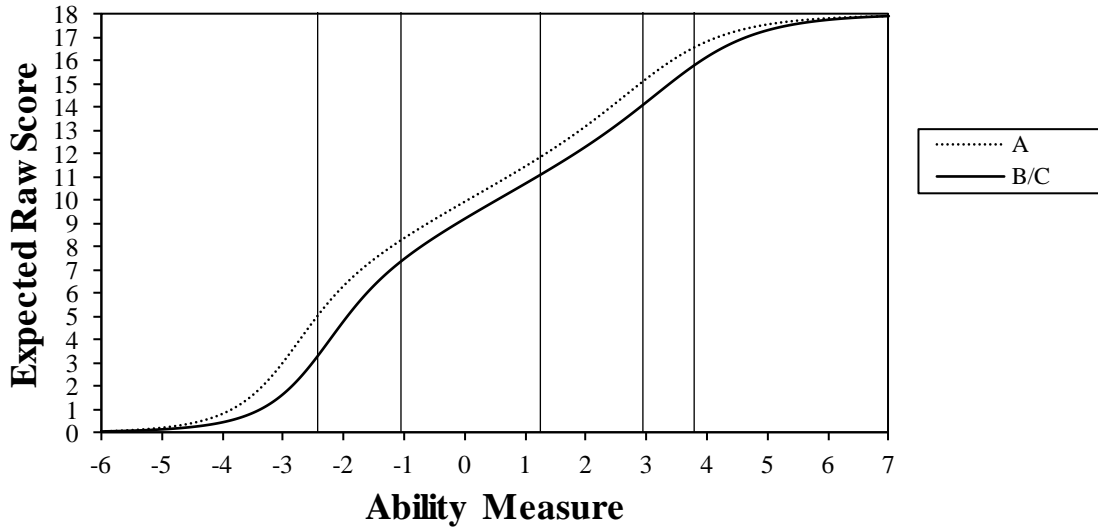


Figure 2.8.3.1.3
 Test Characteristic Curve: Writ 1 S502 Online



2.8.3.2 Grades 2–3

Figure 2.8.3.2.1
 Test Characteristic Curve: Writ 2-3 A S502 Online

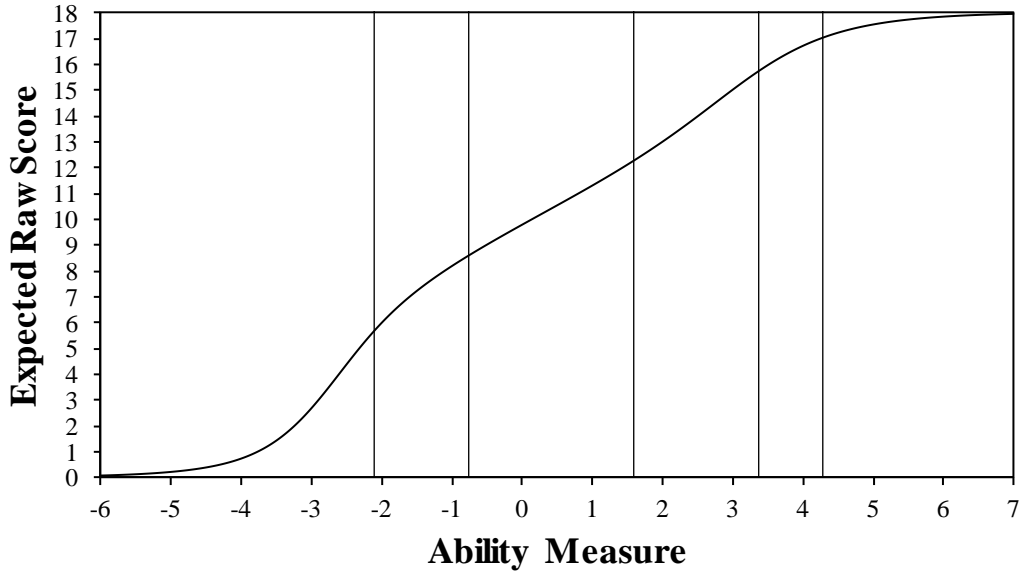


Figure 2.8.3.2.2
 Test Characteristic Curve: Writ 2-3 B/C S502 Online

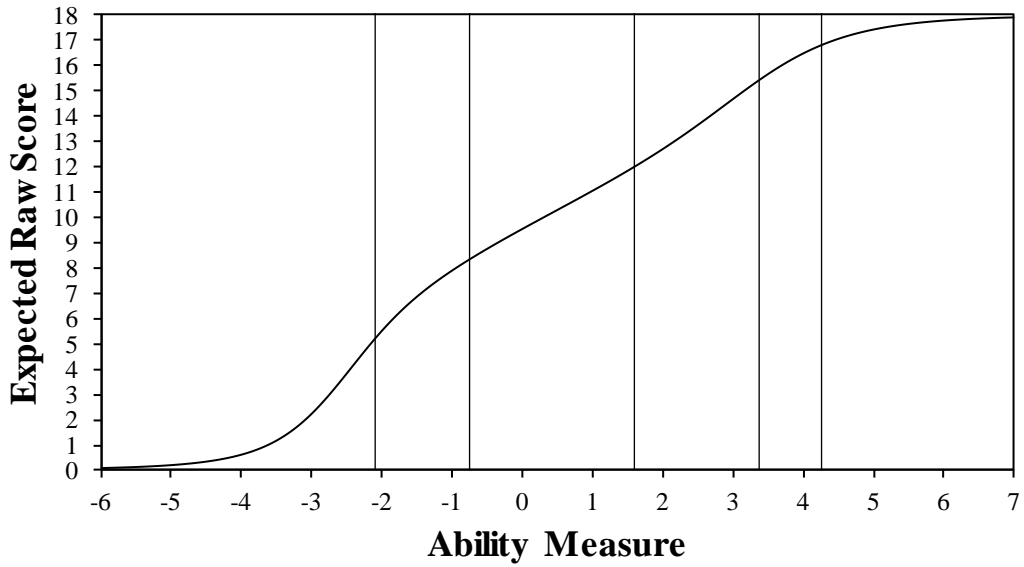
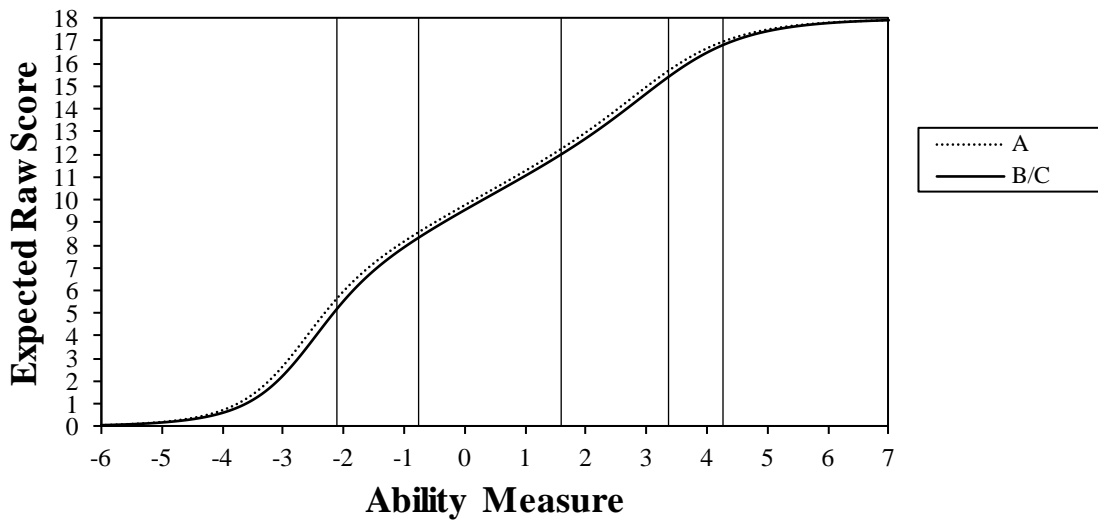


Figure 2.8.3.2.3
 Test Characteristic Curve: Writ 2-3 S502 Online



2.8.3.3 Grades 4–5

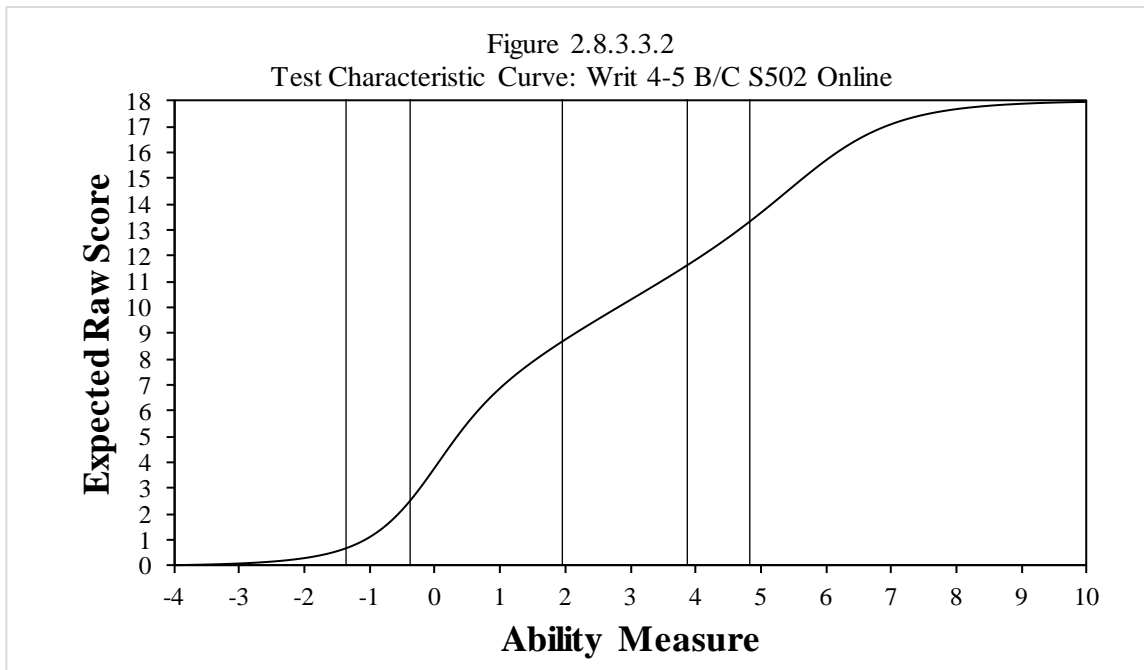
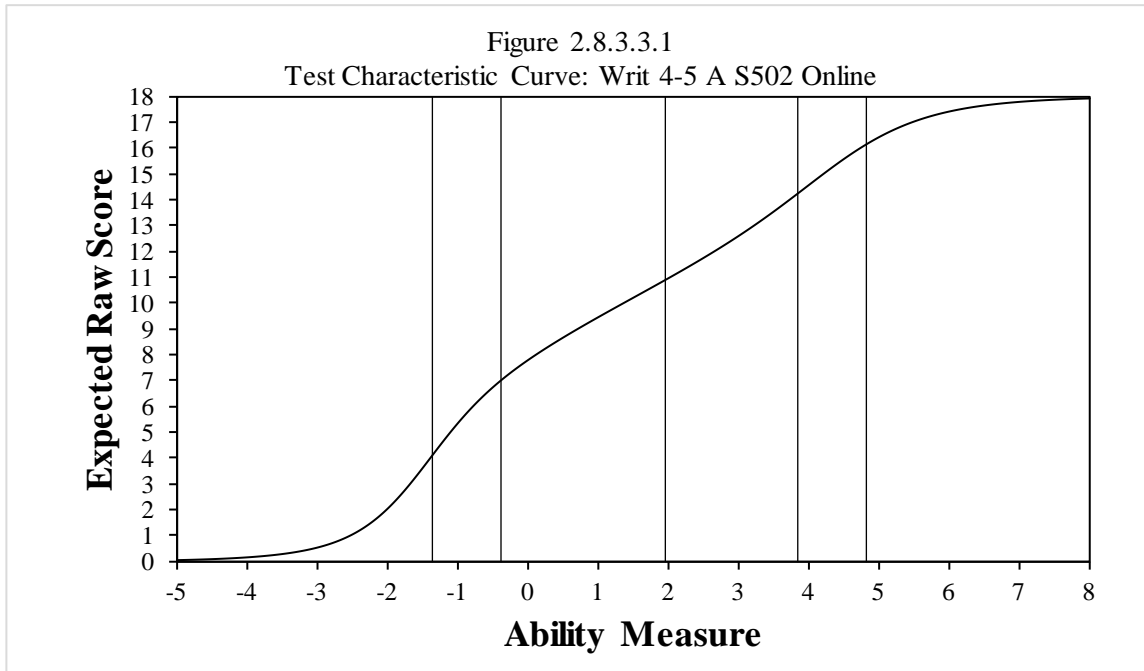
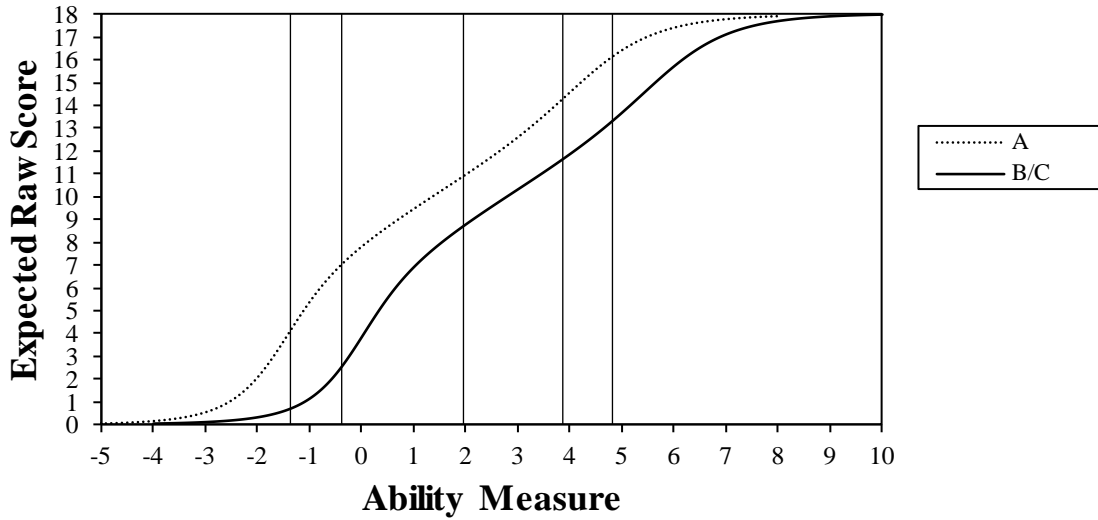


Figure 2.8.3.3
 Test Characteristic Curve: Writ 4-5 S502 Online



2.8.3.4 Grades 6–8

Figure 2.8.3.4.1
 Test Characteristic Curve: Writ 6-8 A S502 Online

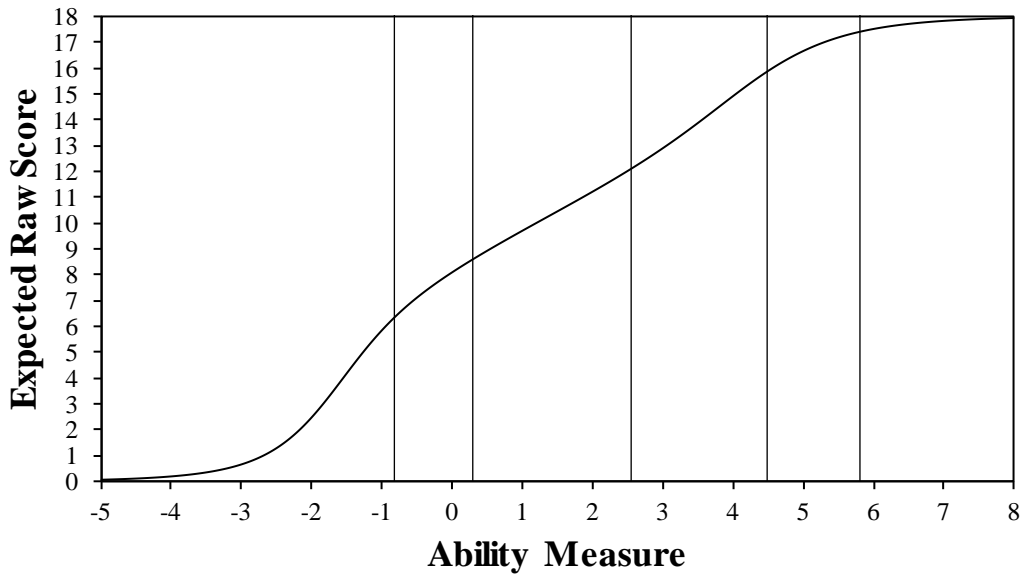


Figure 2.8.3.4.2
 Test Characteristic Curve: Writ 6-8 B/C S502 Online

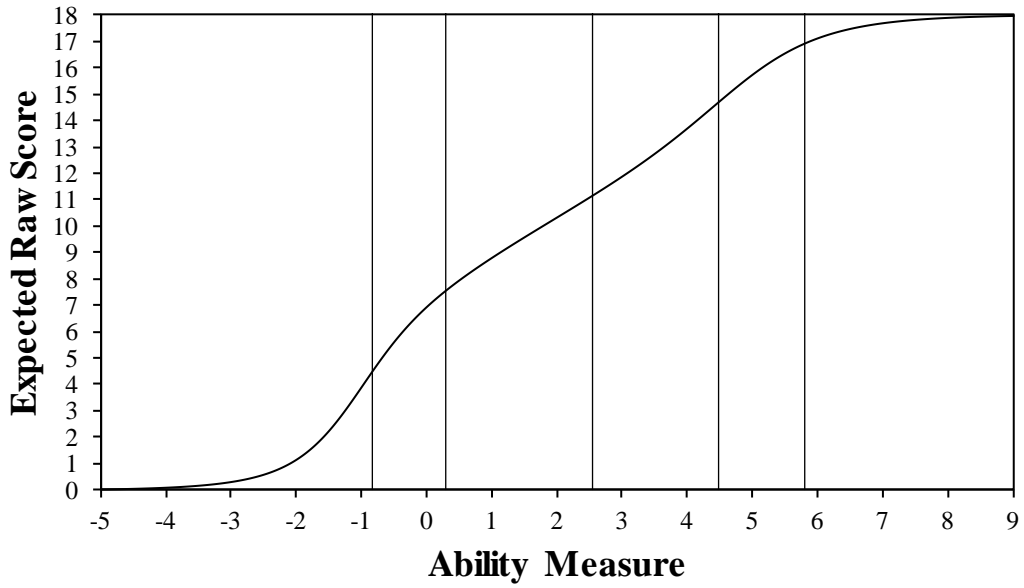


Figure 2.8.3.4.3
 Test Characteristic Curve: Writ 6-8 S502 Online



2.8.3.5 Grades 9-12

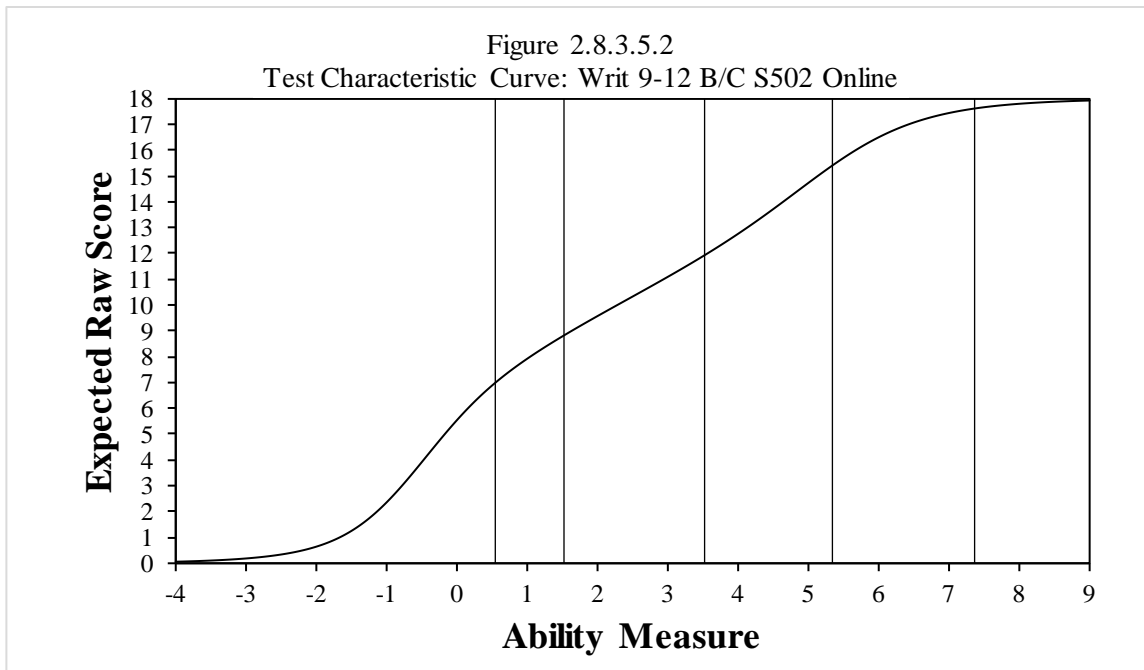
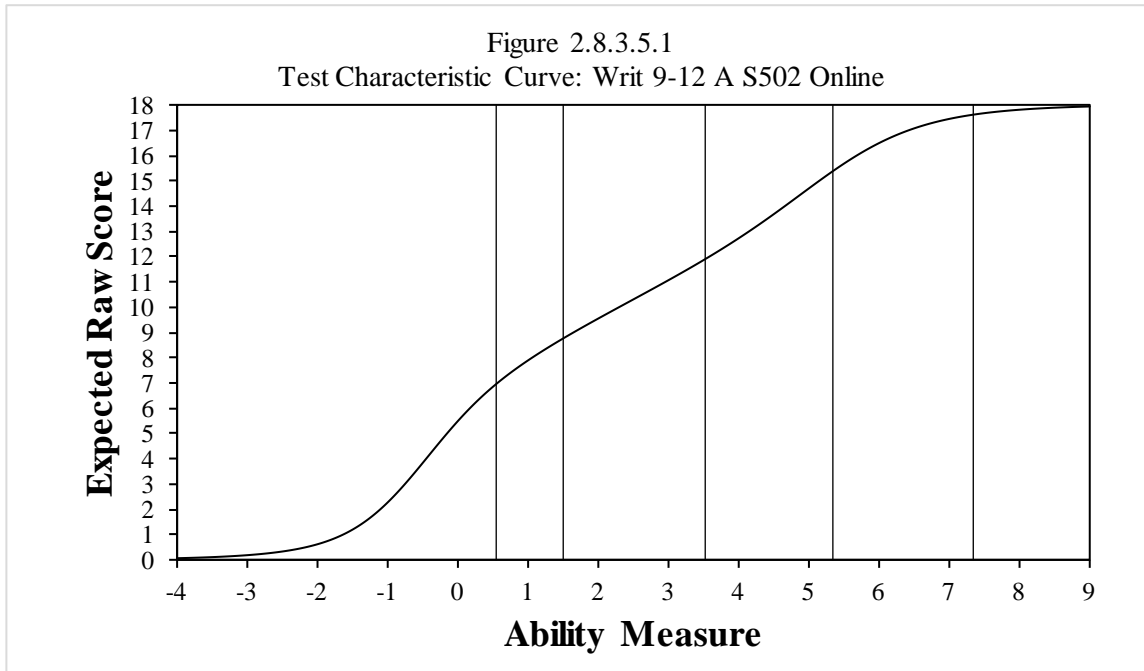
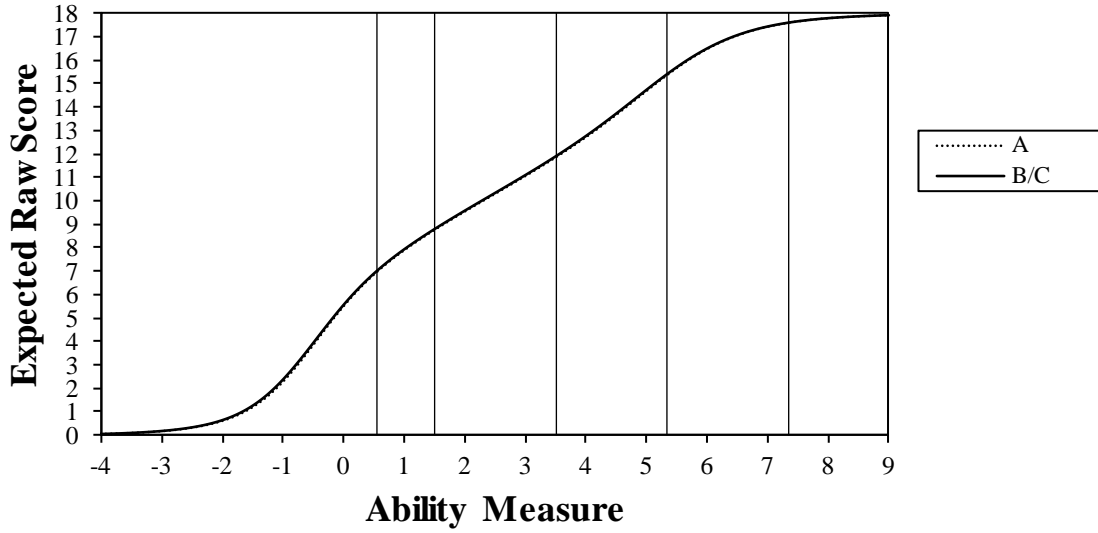


Figure 2.8.3.5.3
 Test Characteristic Curve: Writ 9-12 S502 Online



2.8.4 Speaking

2.8.4.1 Grade 1

Figure 2.8.4.1.1
 Test Characteristic Curve: Spek 1 Pre-A S502 Online

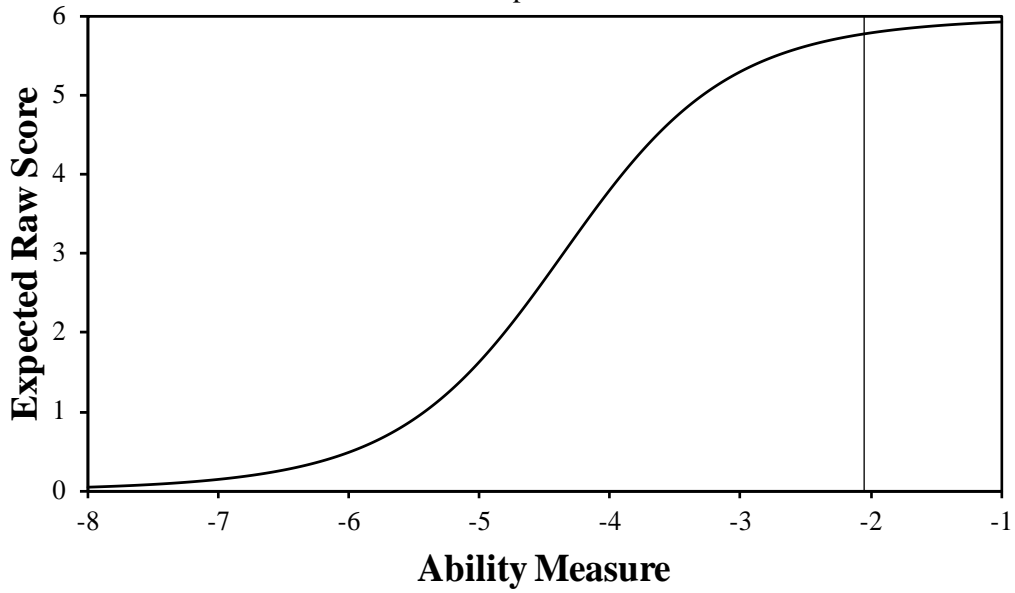


Figure 2.8.4.1.2
Test Characteristic Curve: Spek 1 A S502 Online

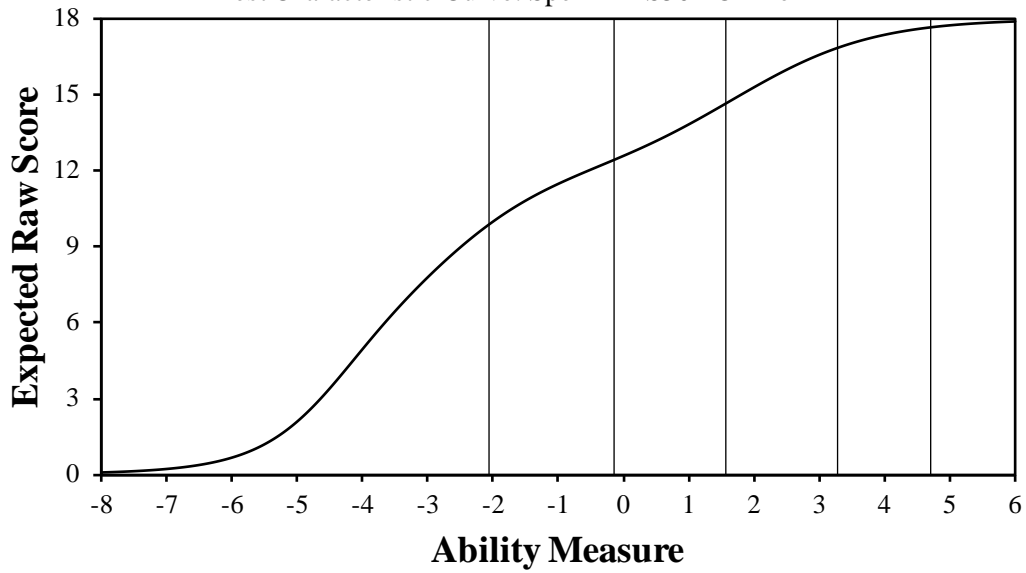


Figure 2.8.4.1.3
Test Characteristic Curve: Spek 1 B/C S502 Online

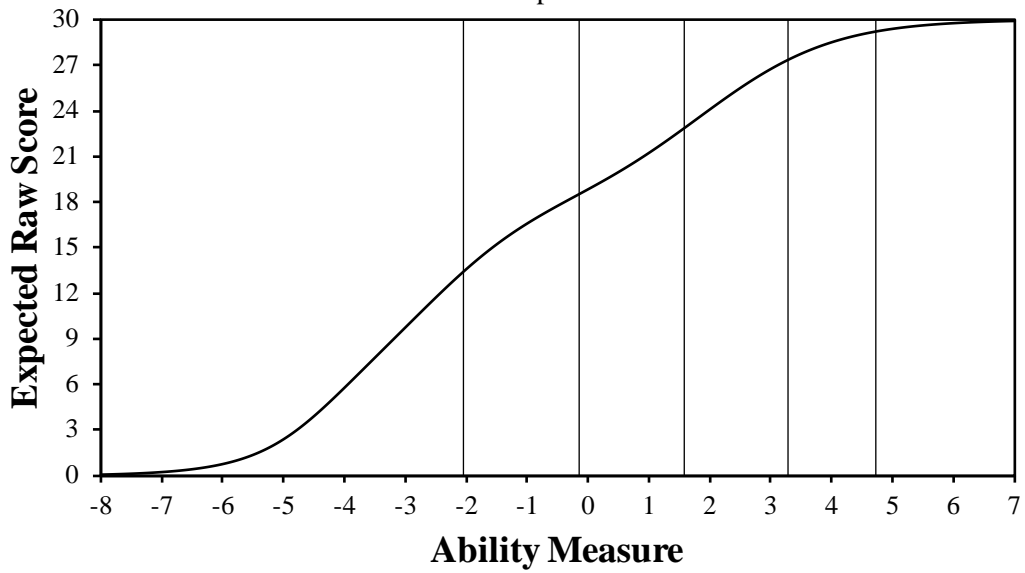
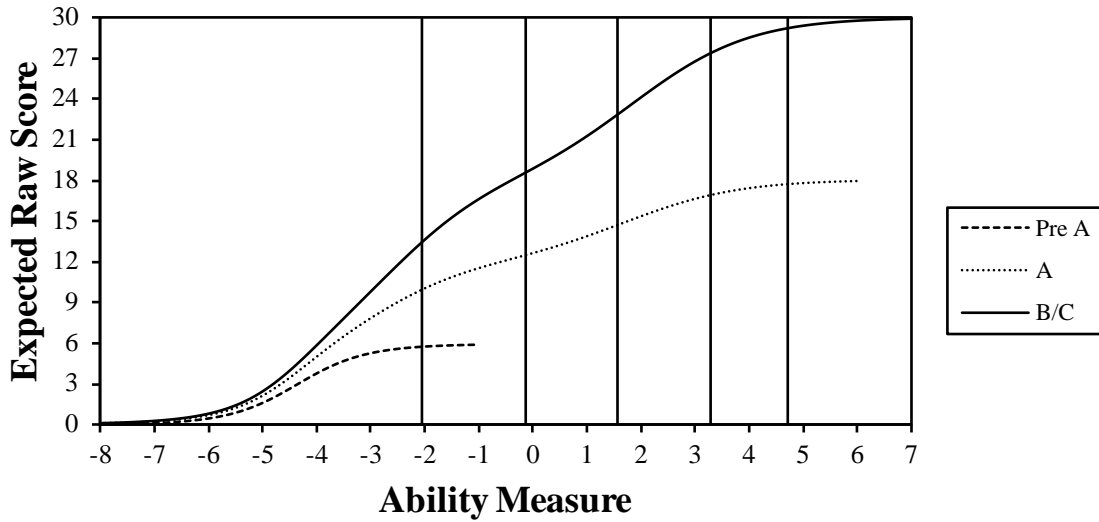


Figure 2.8.4.1.4
 Test Characteristic Curve: Spek 1 S502 Online



2.8.4.2 Grades 2–3

Figure 2.8.4.2.1
 Test Characteristic Curve: Spek 2-3 Pre-A S502 Online

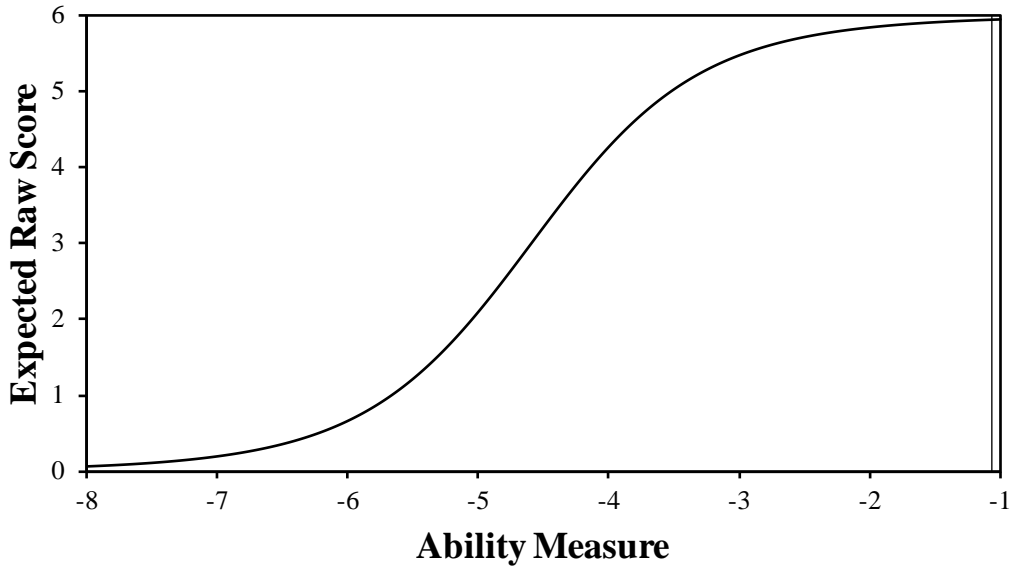


Figure 2.8.4.2.2
Test Characteristic Curve: Spek 2-3 A S502 Online

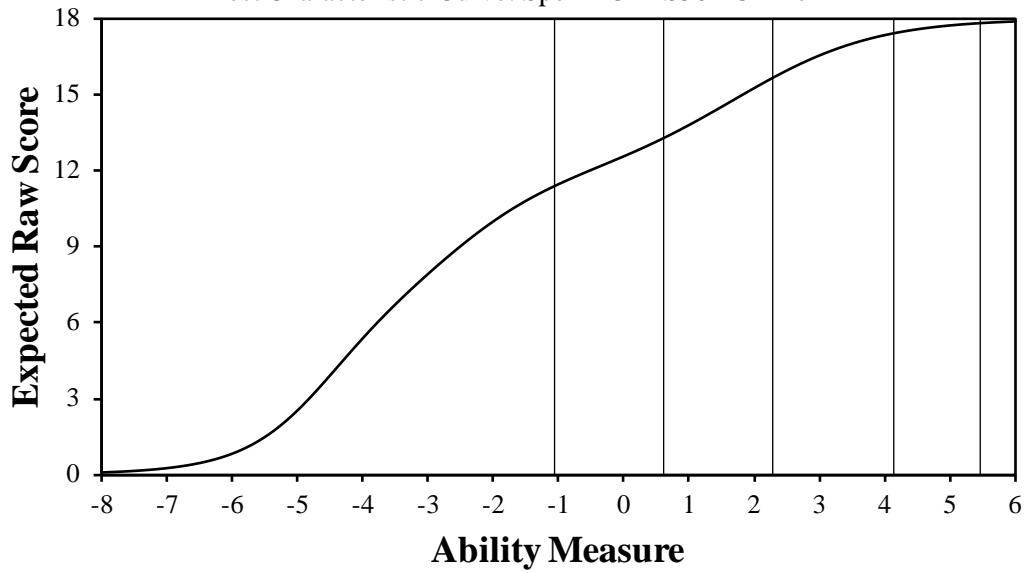


Figure 2.8.4.2.3
Test Characteristic Curve: Spek 2-3 B/C S502 Online

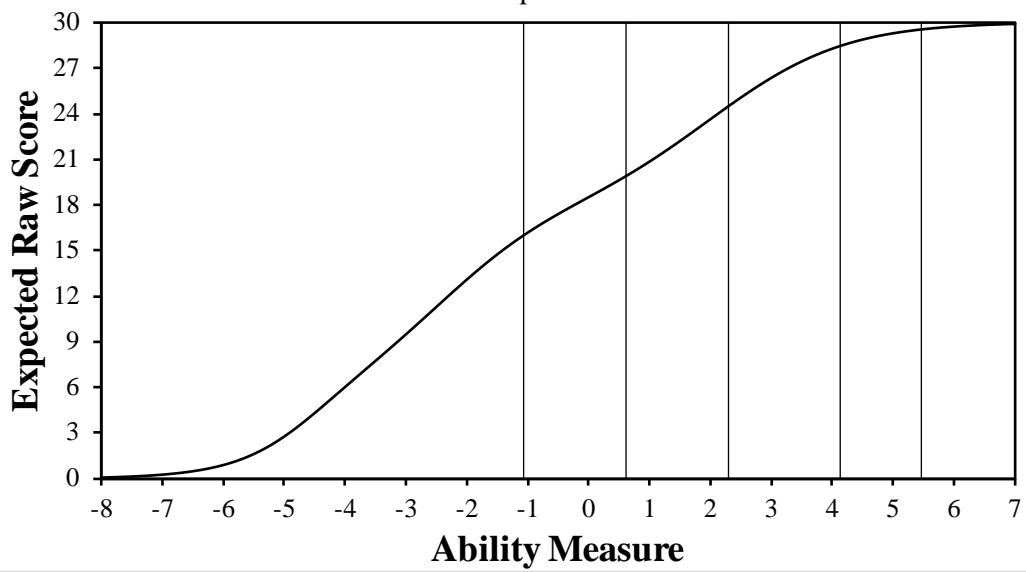
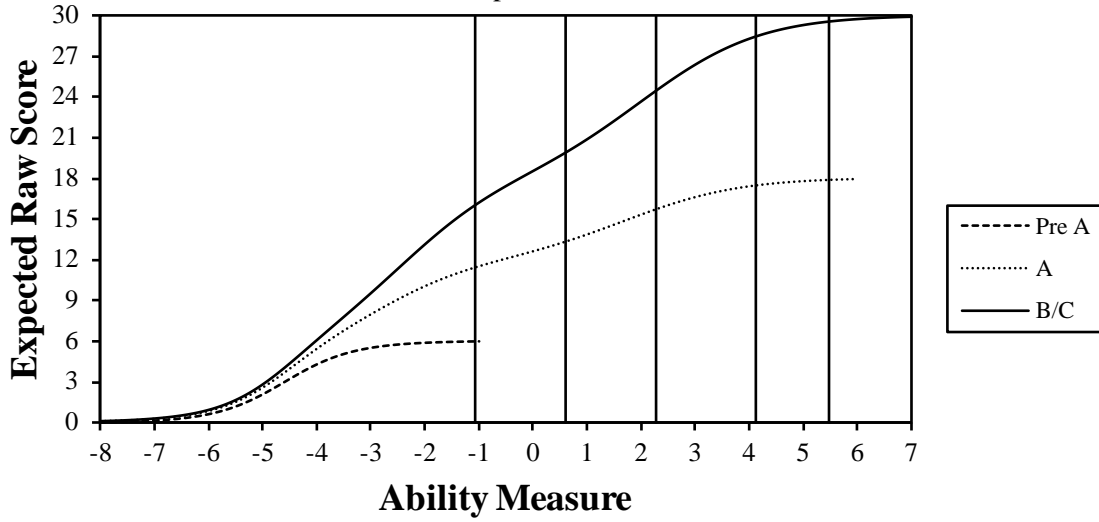


Figure 2.8.4.2.4
 Test Characteristic Curve: Spek 2-3 S502 Online



2.8.4.3 Grades 4–5

Figure 2.8.4.3.1
 Test Characteristic Curve: Spek 4-5 Pre-A S502 Online

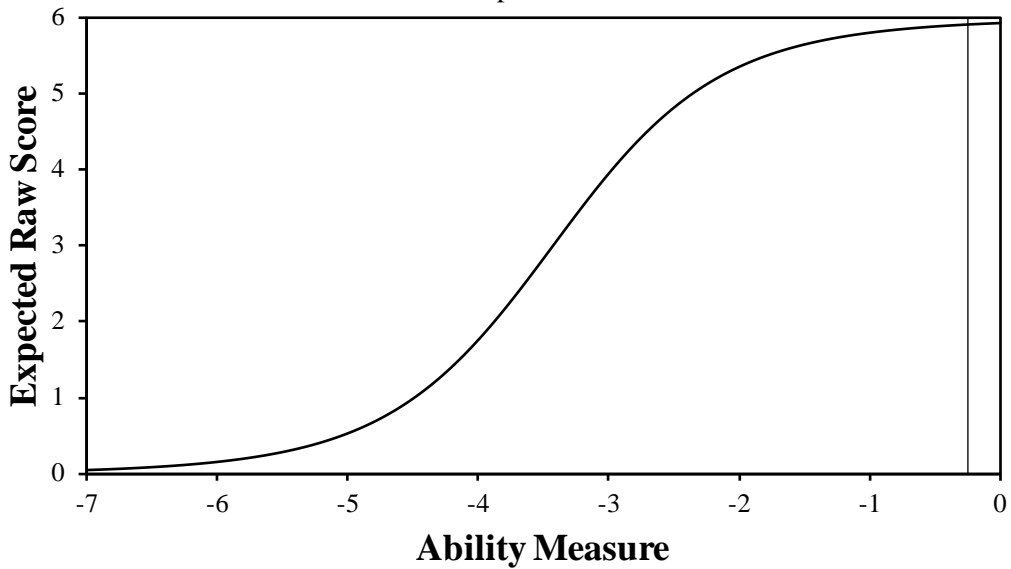


Figure 2.8.4.3.2
Test Characteristic Curve: Spek 4-5 A S502 Online

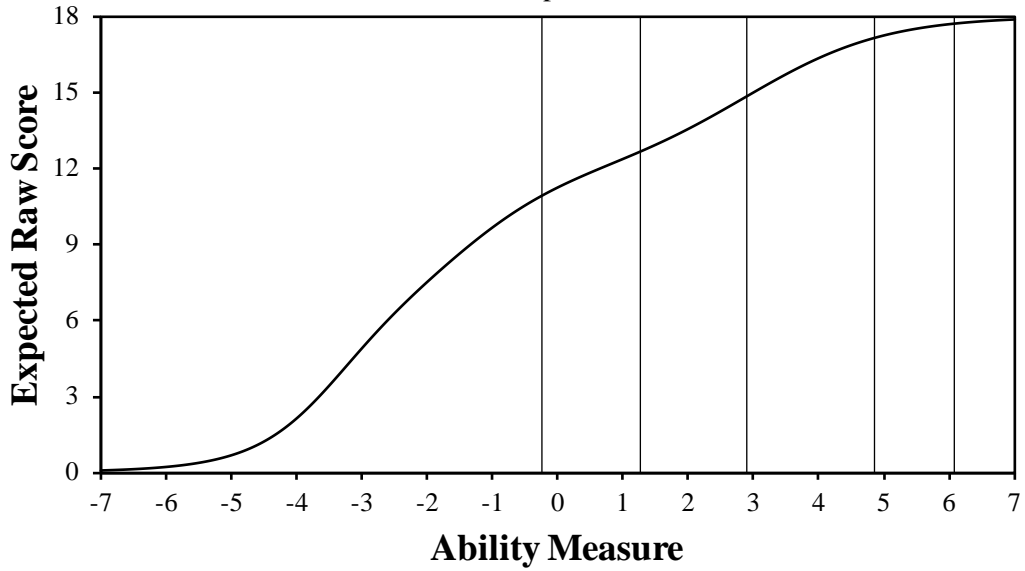


Figure 2.8.4.3.3
Test Characteristic Curve: Spek 4-5 B/C S502 Online

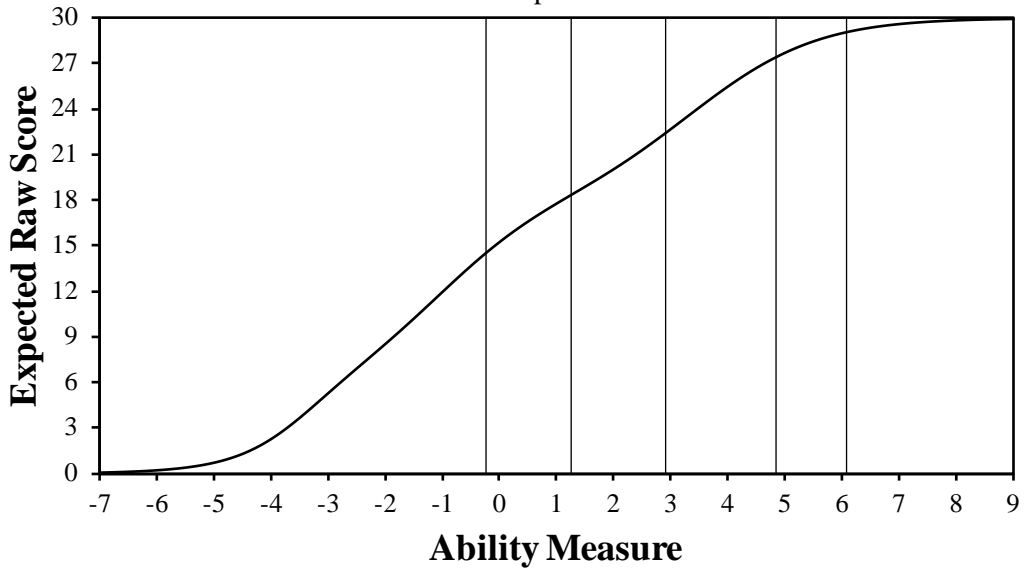
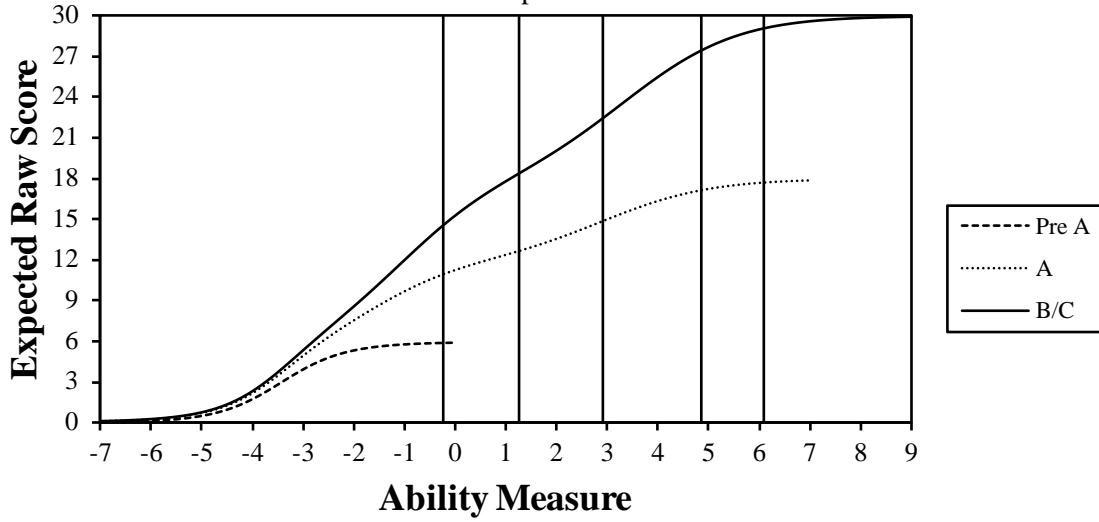


Figure 2.8.4.3.4
 Test Characteristic Curve: Spek 4-5 S502 Online



2.8.4.4 Grades 6–8

Figure 2.8.4.4.1
 Test Characteristic Curve: Spek 6-8 Pre-A S502 Online

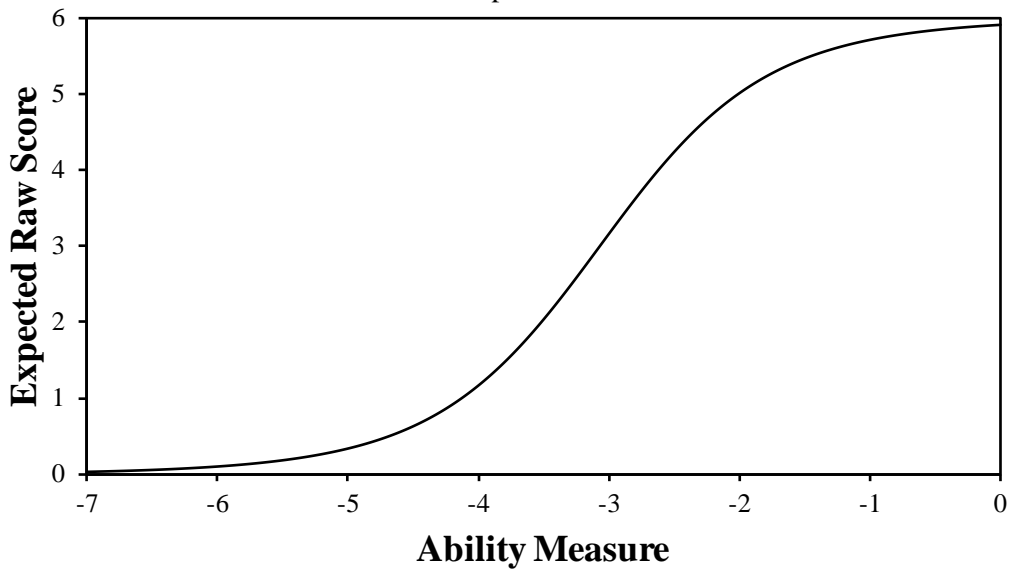


Figure 2.8.4.4.2
Test Characteristic Curve: Spek 6-8 A S502 Online

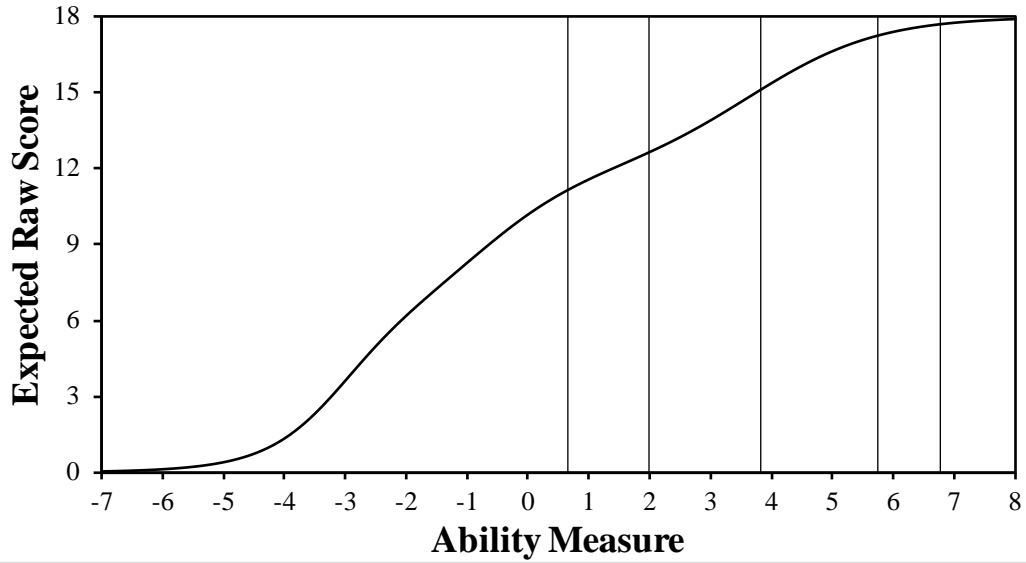


Figure 2.8.4.4.3
Test Characteristic Curve: Spek 6-8 B/C S502 Online

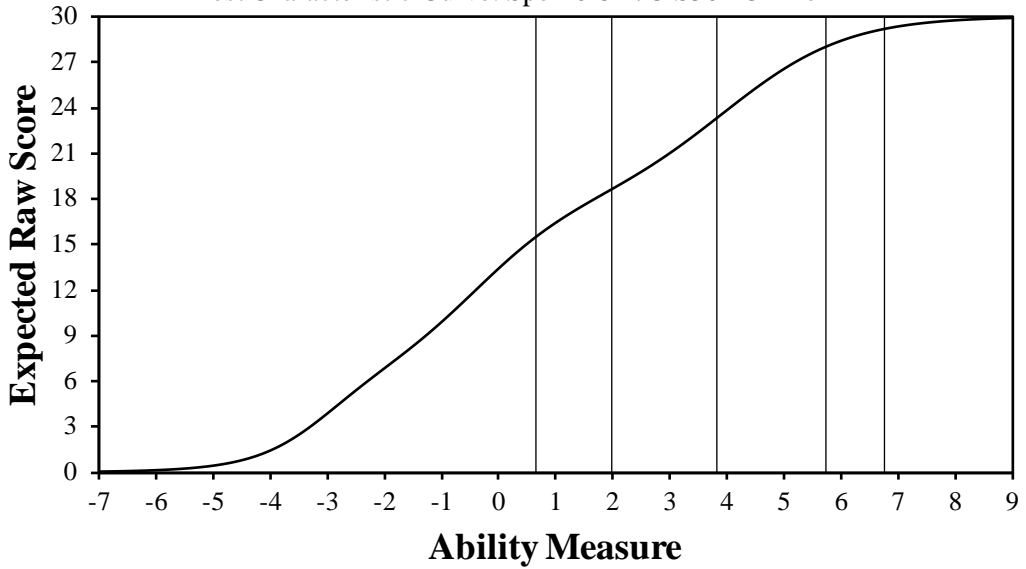
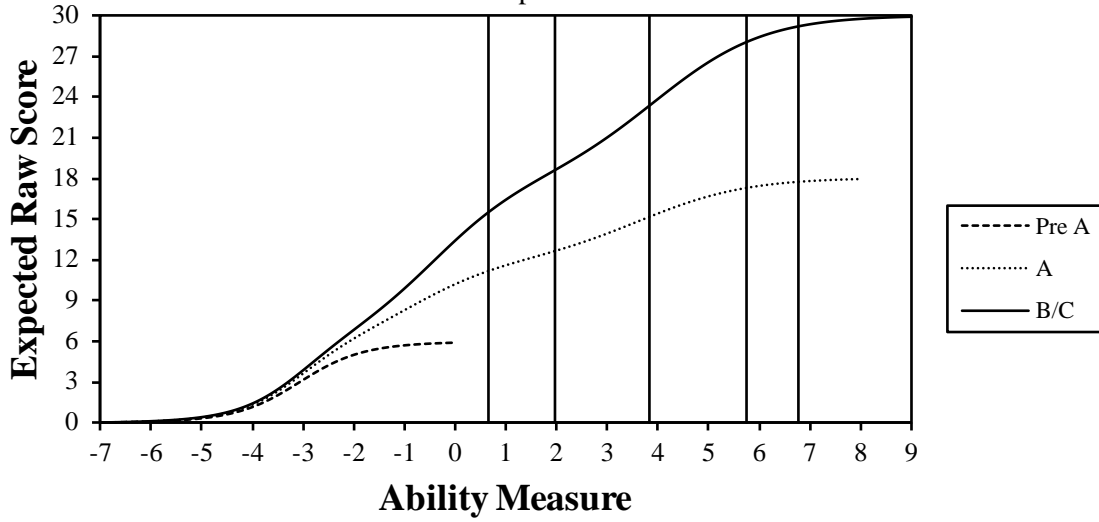
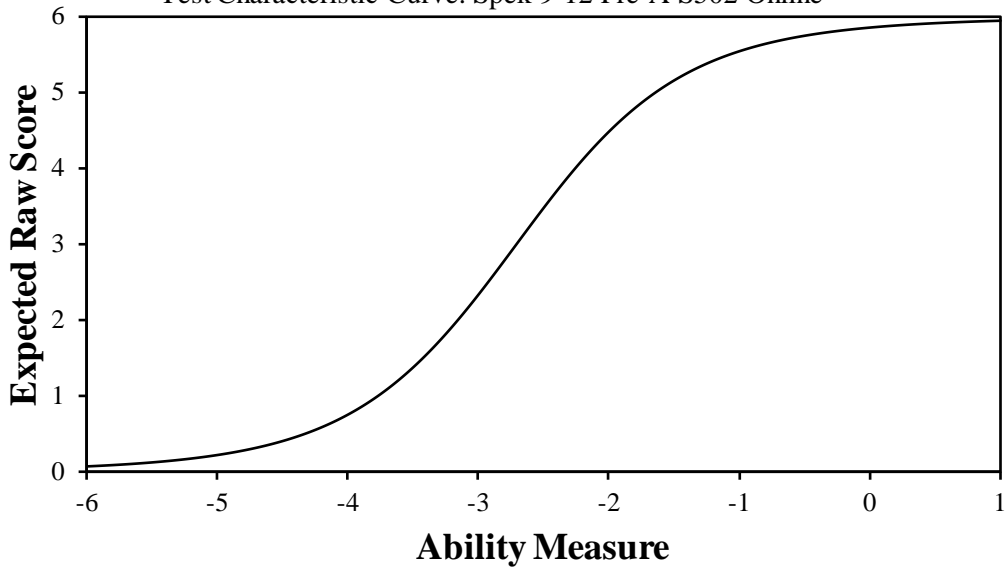


Figure 2.8.4.4
 Test Characteristic Curve: Spek 6-8 S502 Online



2.8.4.5 Grades 9-12

Figure 2.8.4.5.1
 Test Characteristic Curve: Spek 9-12 Pre-A S502 Online



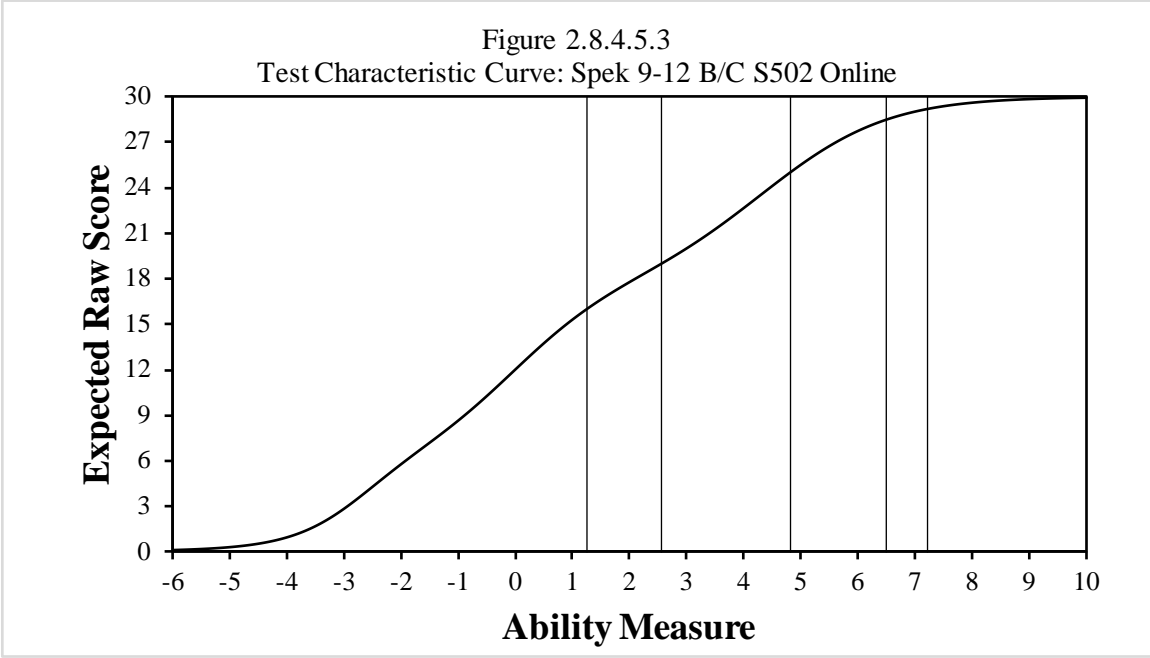
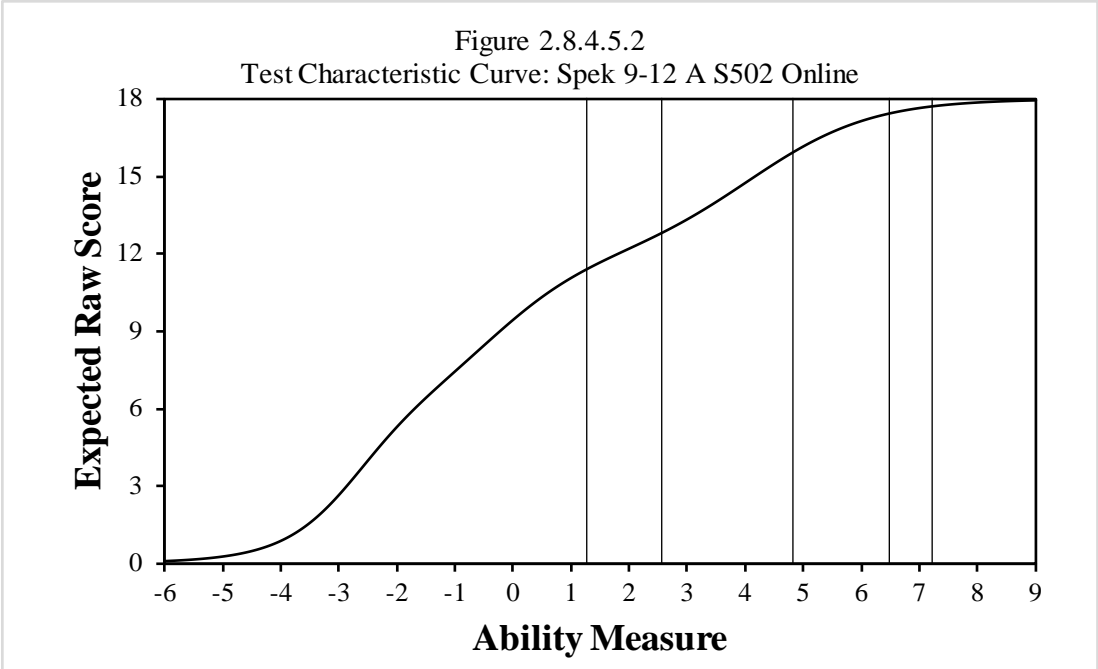
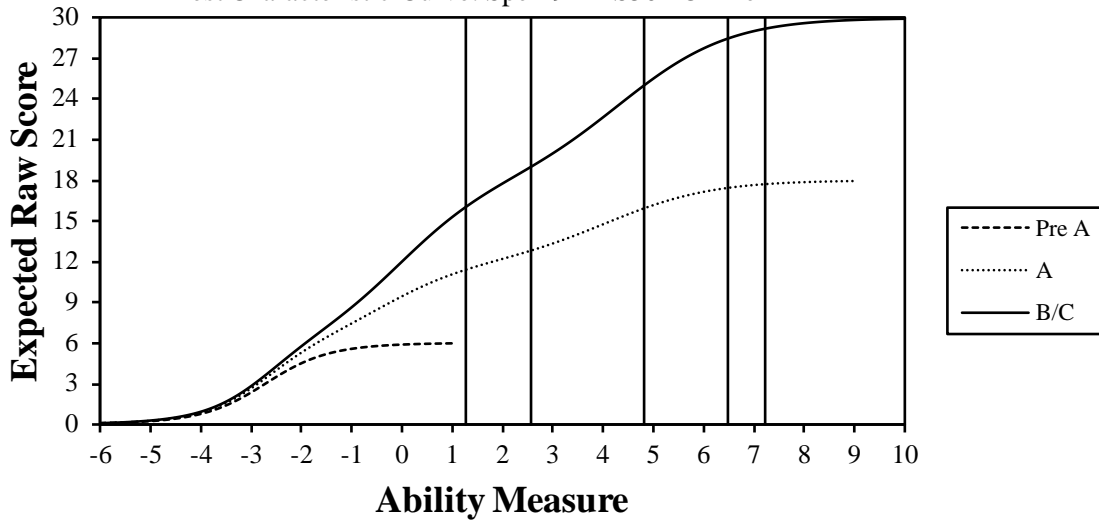


Figure 2.8.4.5.4
Test Characteristic Curve: Spek 9-12 S502 Online



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 on 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 rating 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 rating 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 scale to assign scores (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).

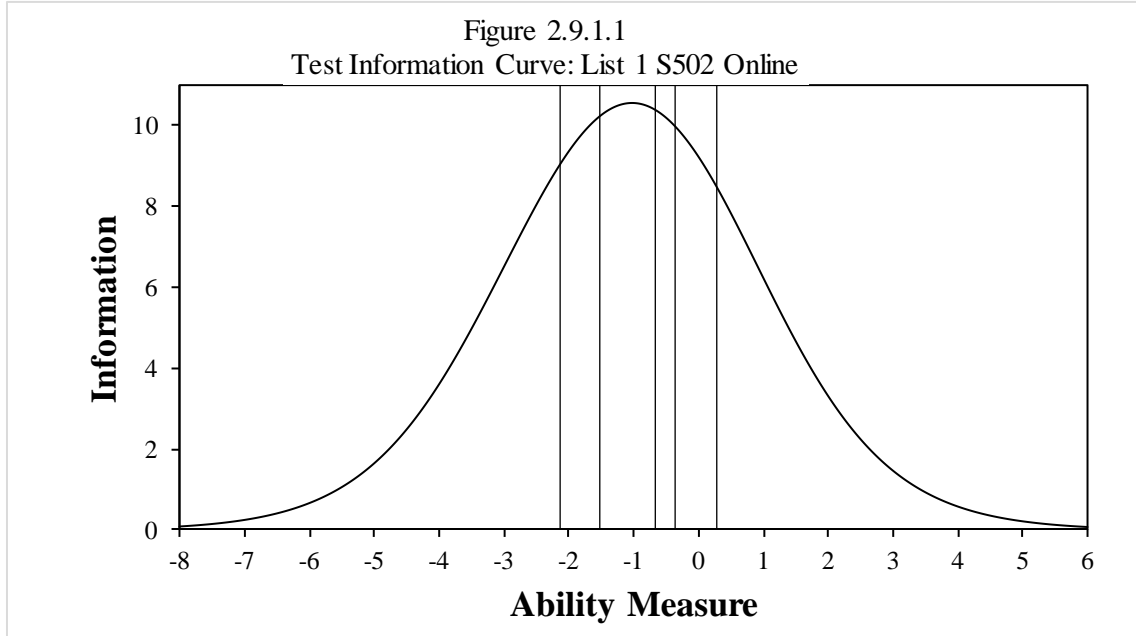
The figures in this section plot the TIFs and show graphically the amount of information that the test provided across the continuum of student ability. For each test form, the five vertical lines in the figure indicate the ACCESS cut scores for the highest grade in each grade-level cluster, dividing the figure into six sections denoting the WIDA proficiency levels (1–6) 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 is 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 students' ability. Readers should note that WIDA states do not make reclassification decisions based solely on students' domain scale score. Most WIDA states set their reclassification or exit criterion based on students' Overall Composite scale score. Students' Overall Composite scale score is a weighted sum of the four domain scale scores. Only a few states set their reclassification criterion using both one or more of the ACCESS Composites as well as the individual domain scale scores. Therefore, from the WIDA policy perspective, it is more important to ensure that we minimize the measurement error near the cut point where most states set their reclassification criterion on the Overall Composite scale score. We report the CSEM for ACCESS composites in Section 5.6.

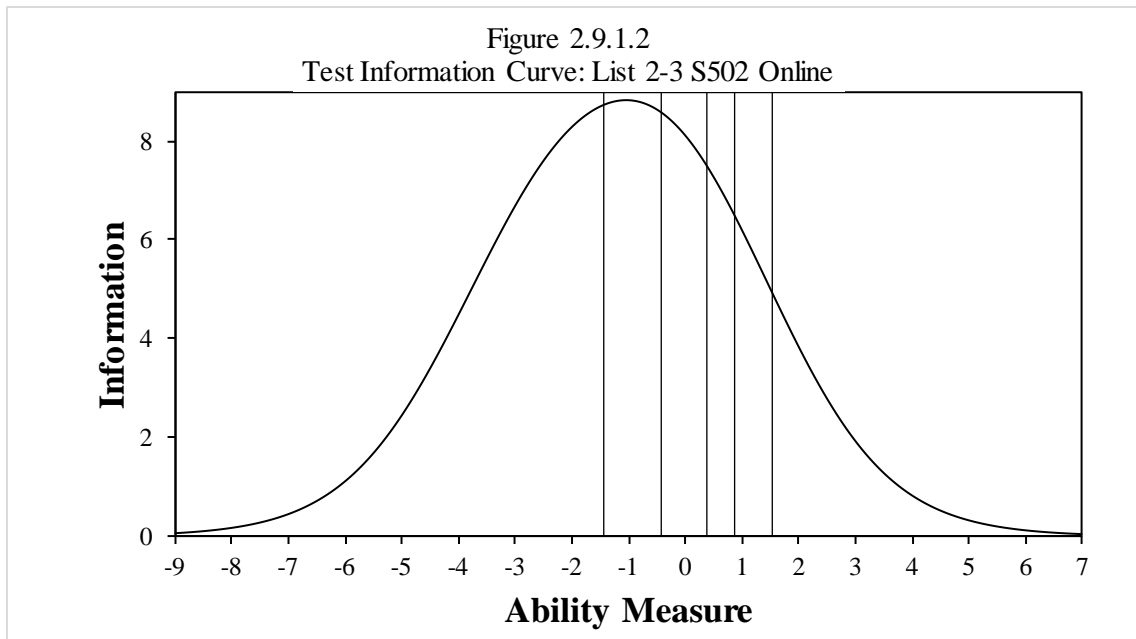
In addition to the TIF graphs by tier, for the Writing and Speaking tests, in the same graph we provide plots of the TIFs across tiers, by grade-level cluster. Test users may find it useful to compare the ability ranges across tiers where the curves display a peak (i.e., where the best measurement information 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 just below the PL 2 cut score, as well as for those student ability measures that were just below the PL 4 cut score. By contrast, the Writing Grade 1 Tier B/C form provided more information about the student ability measures that were just above the PL 2 cut score, and just above the PL 4 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.

2.9.1 Listening

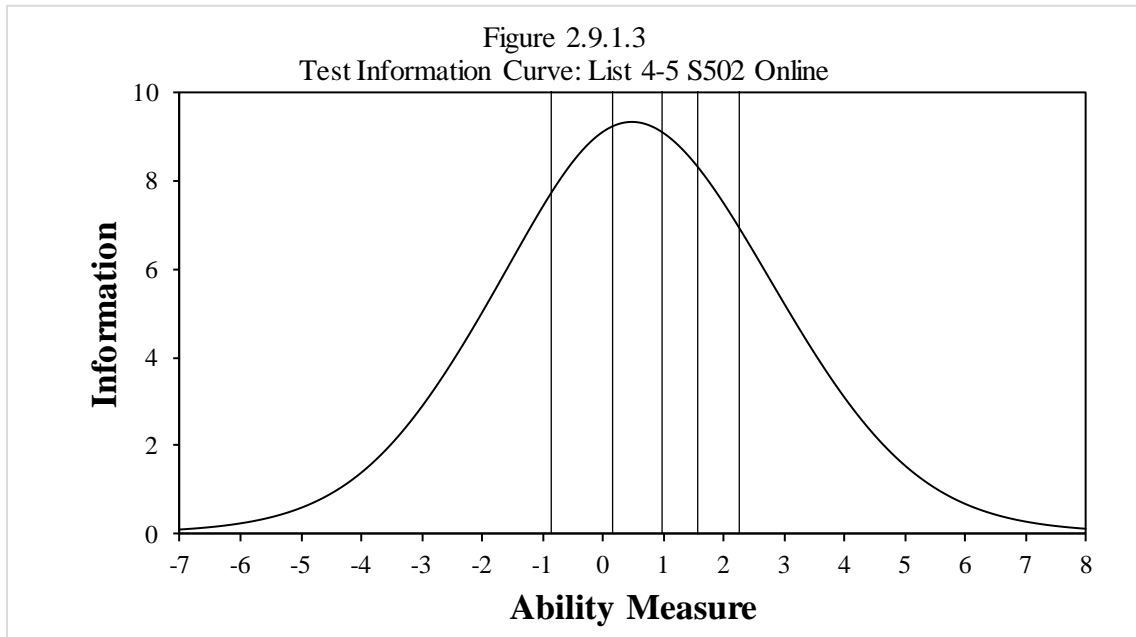
2.9.1.1 Grade 1



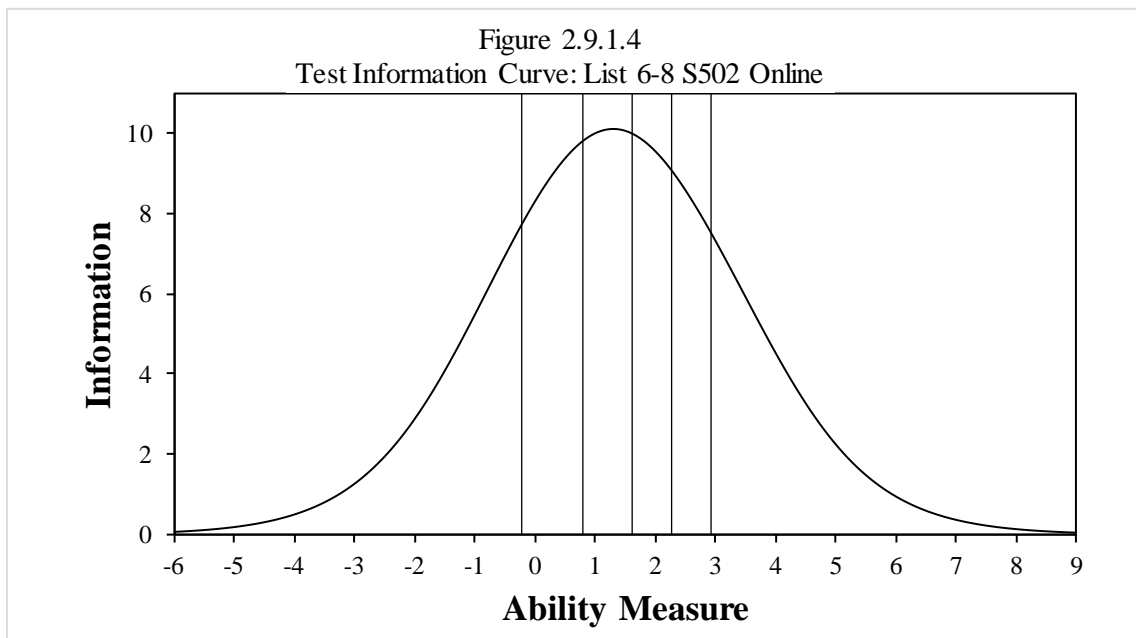
2.9.1.2 Grades 2–3



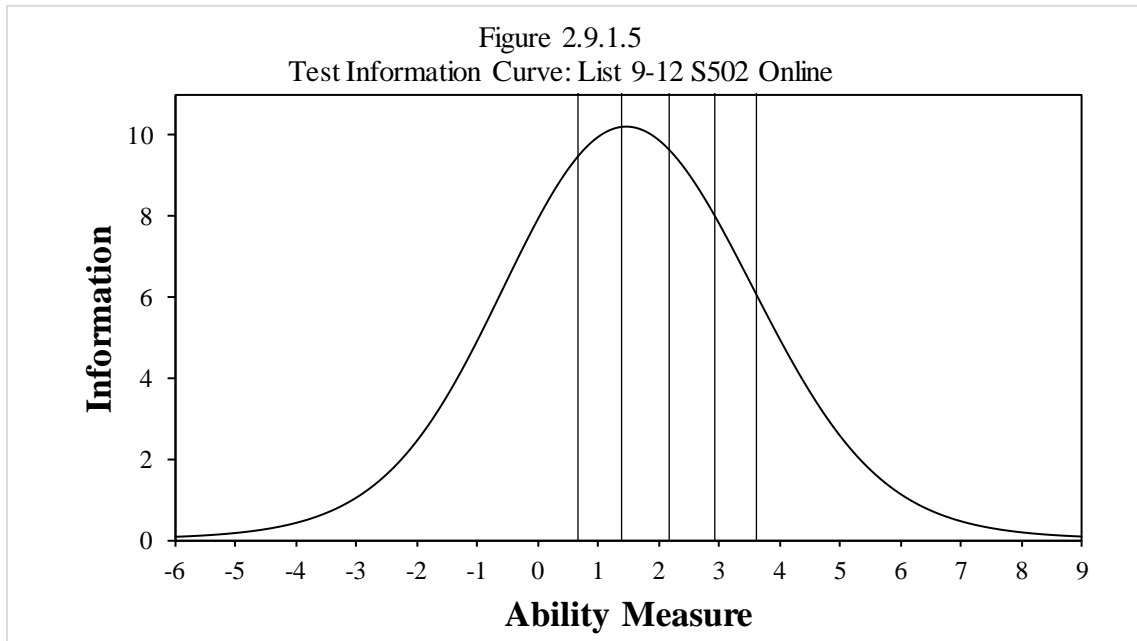
2.9.1.3 Grades 4–5



2.9.1.4 Grades 6–8

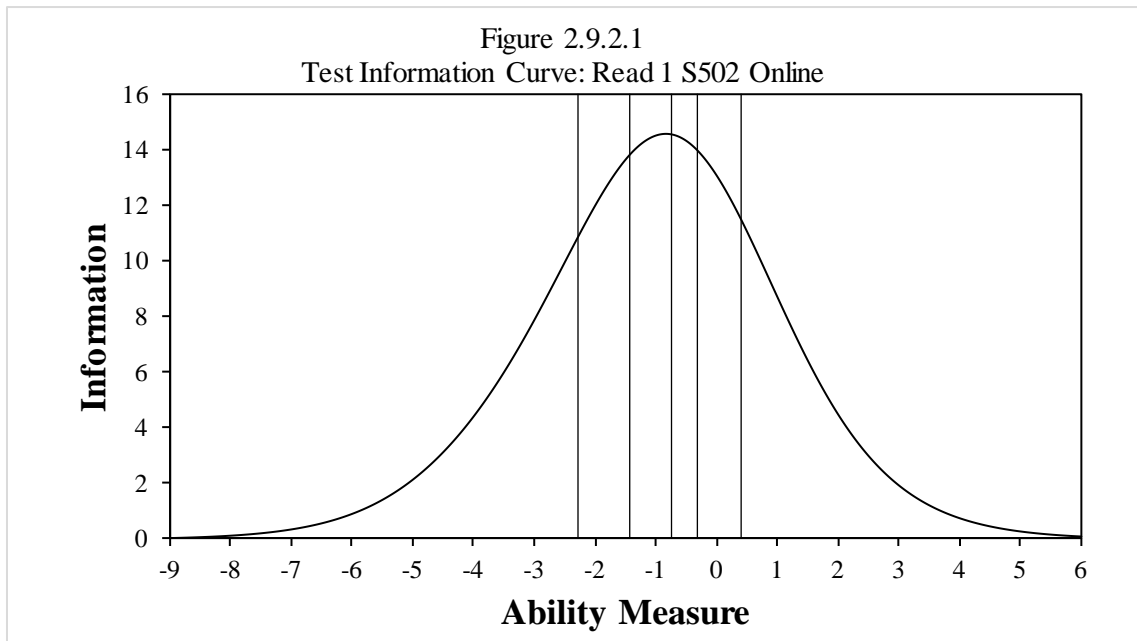


2.9.1.5 Grades 9-12

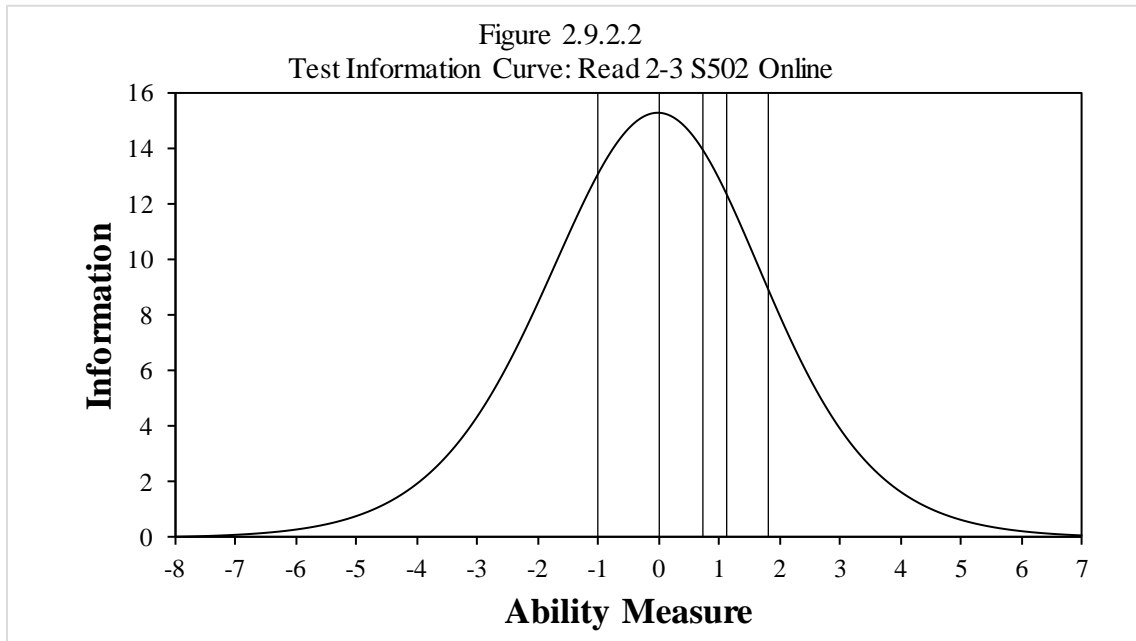


2.9.2 Reading

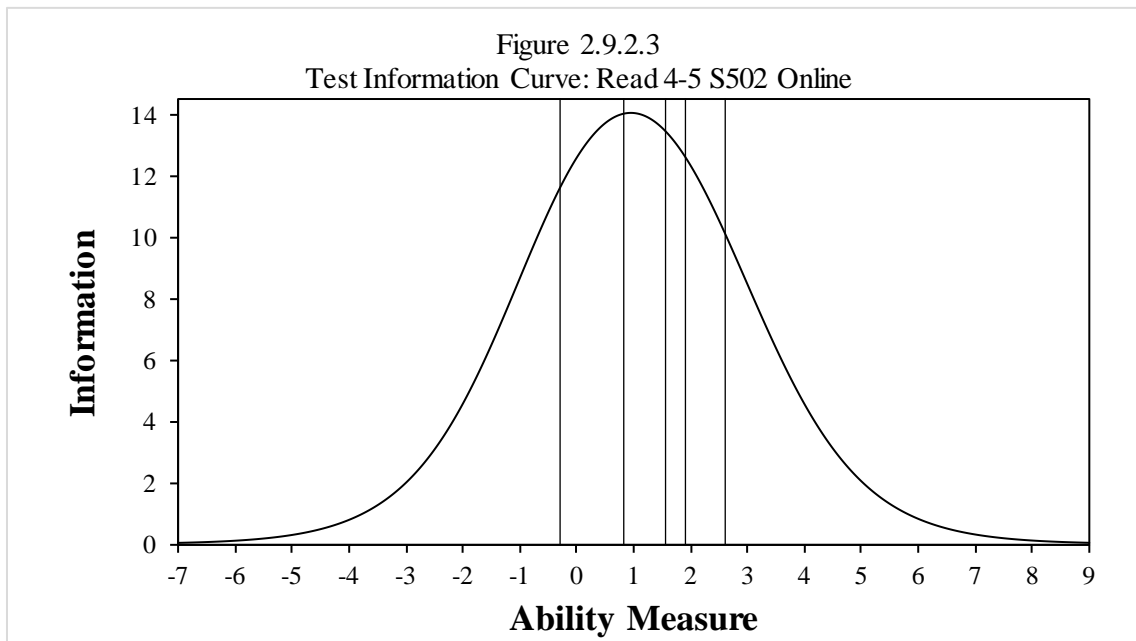
2.9.2.1 Grade 1



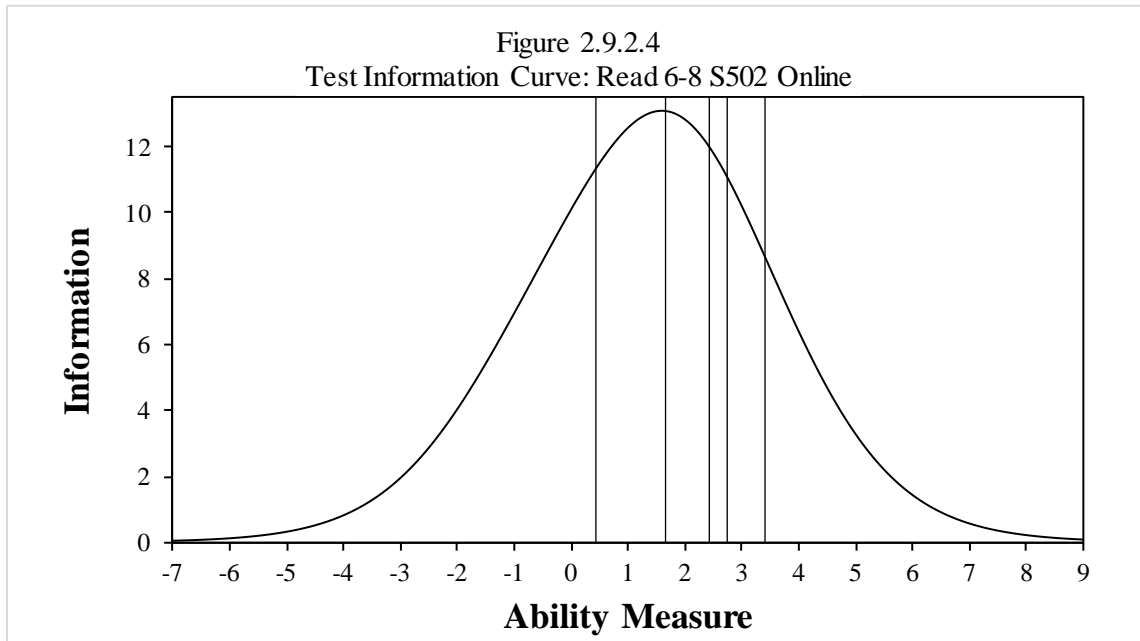
2.9.2.2 Grades 2–3



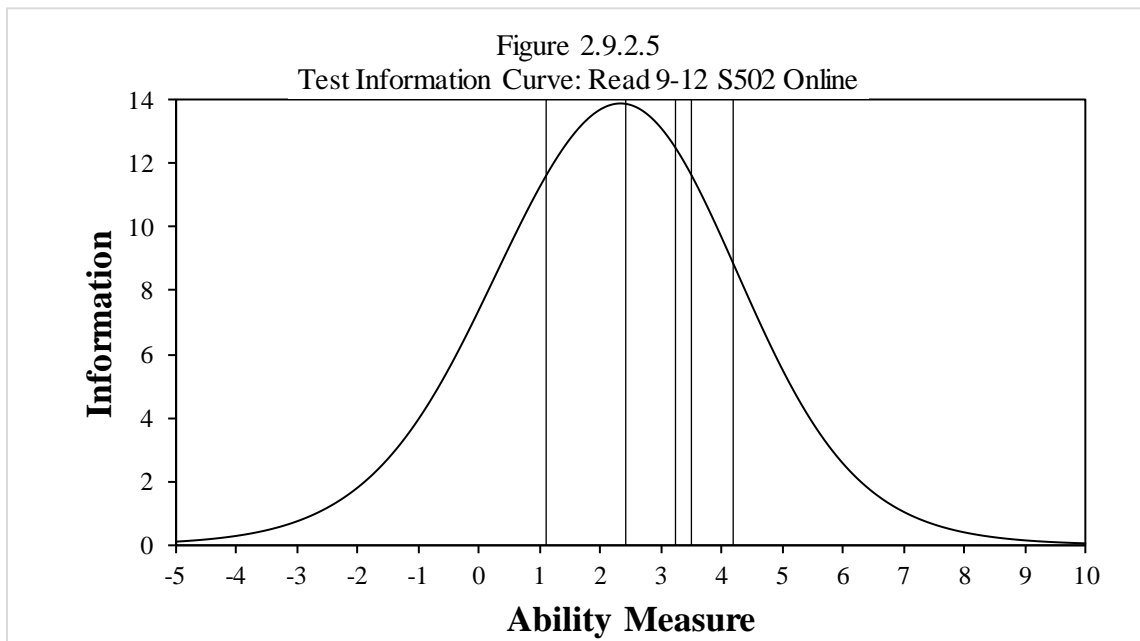
2.9.2.3 Grades 4–5



2.9.2.4 Grades 6–8

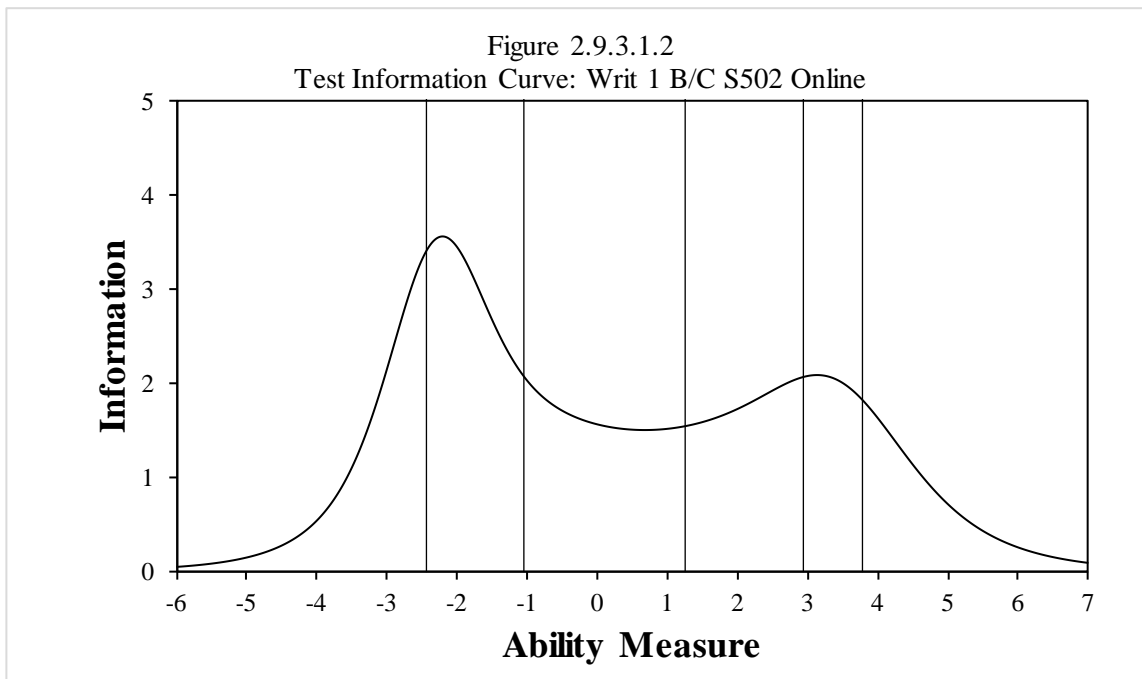
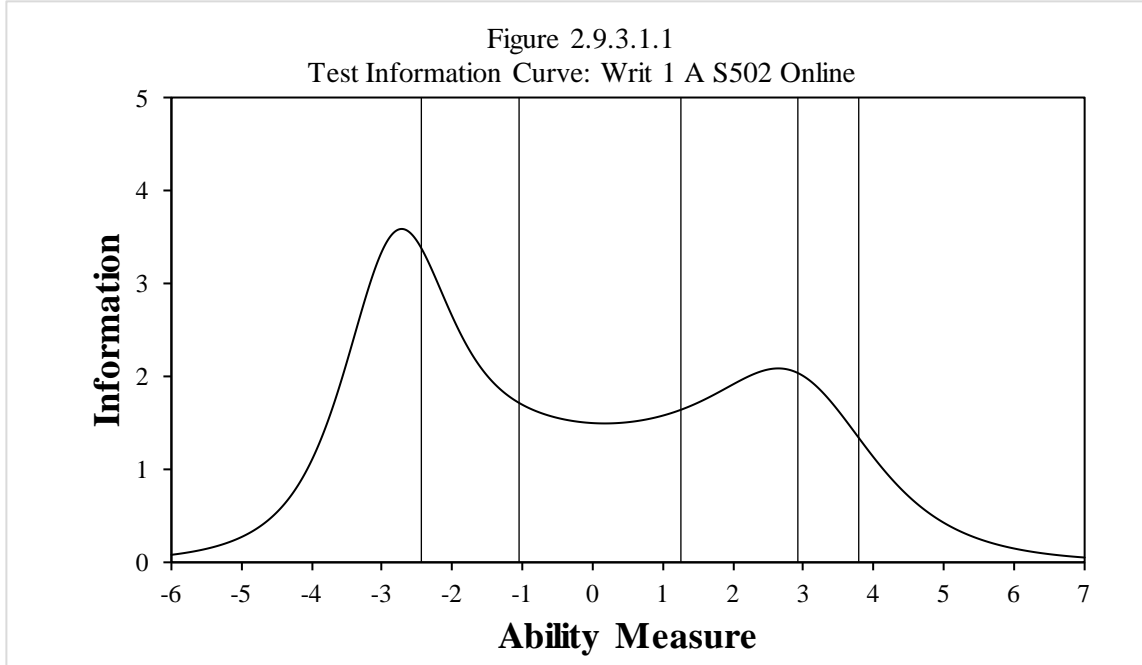


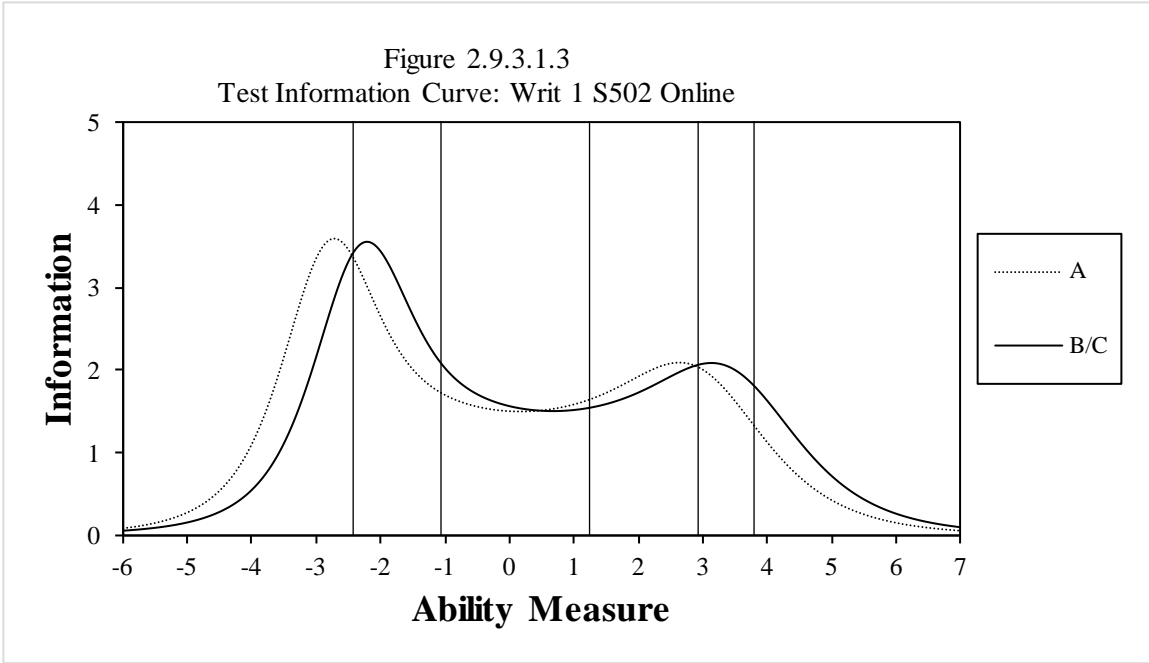
2.9.2.5 Grades 9-12



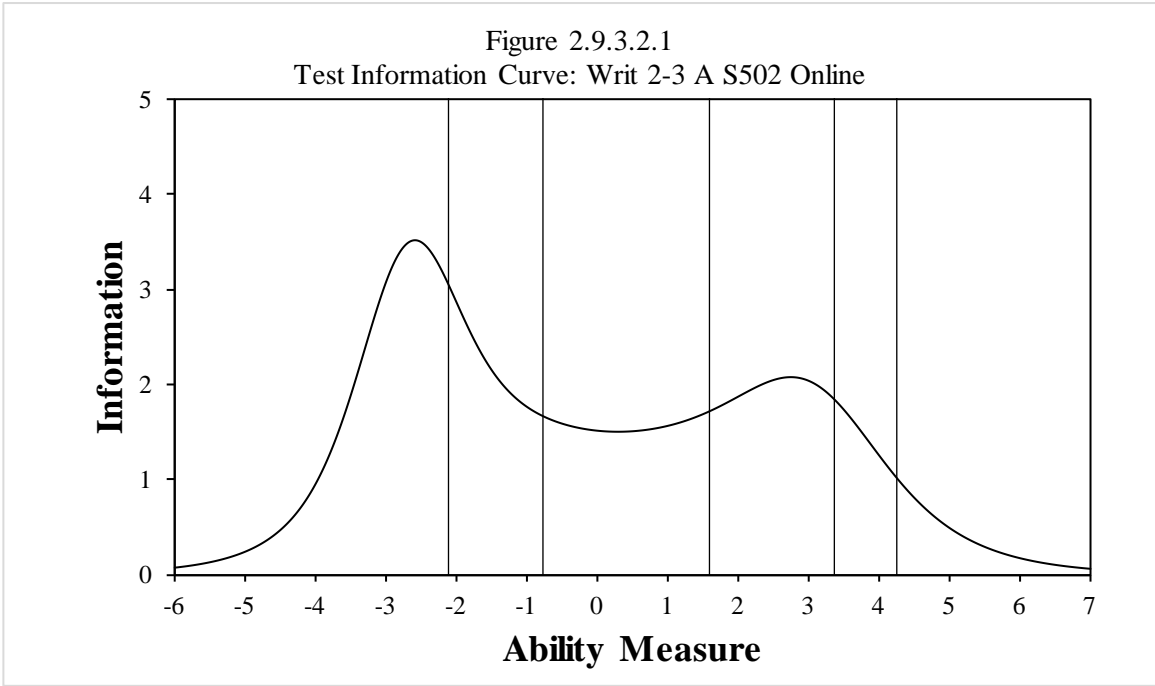
2.9.3 Writing

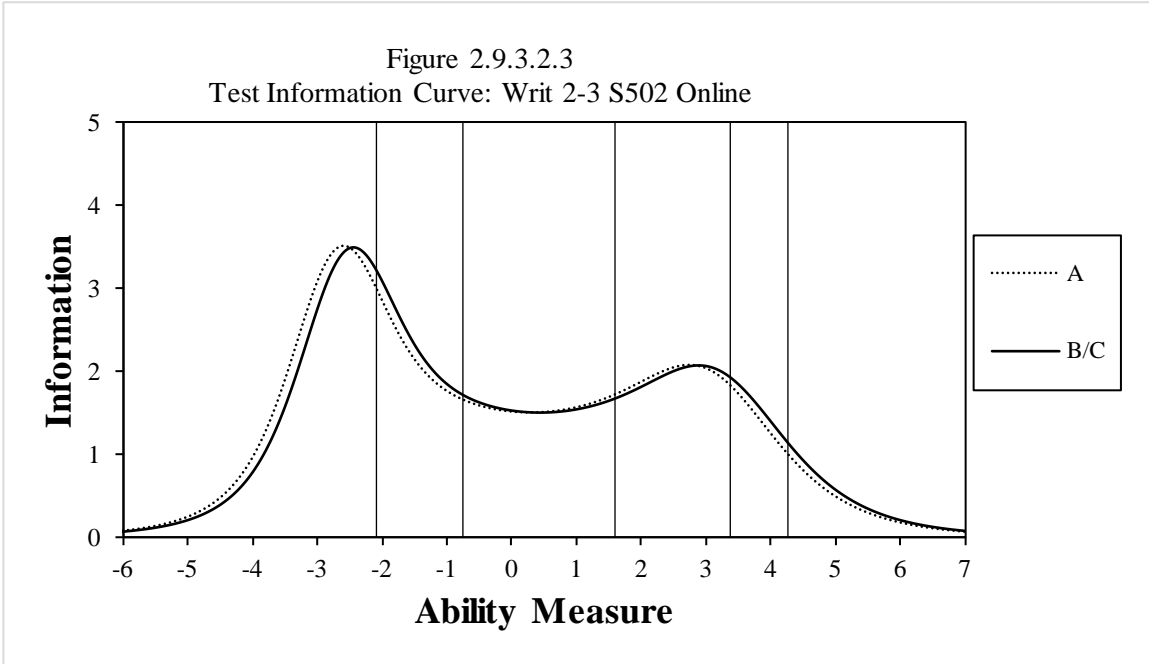
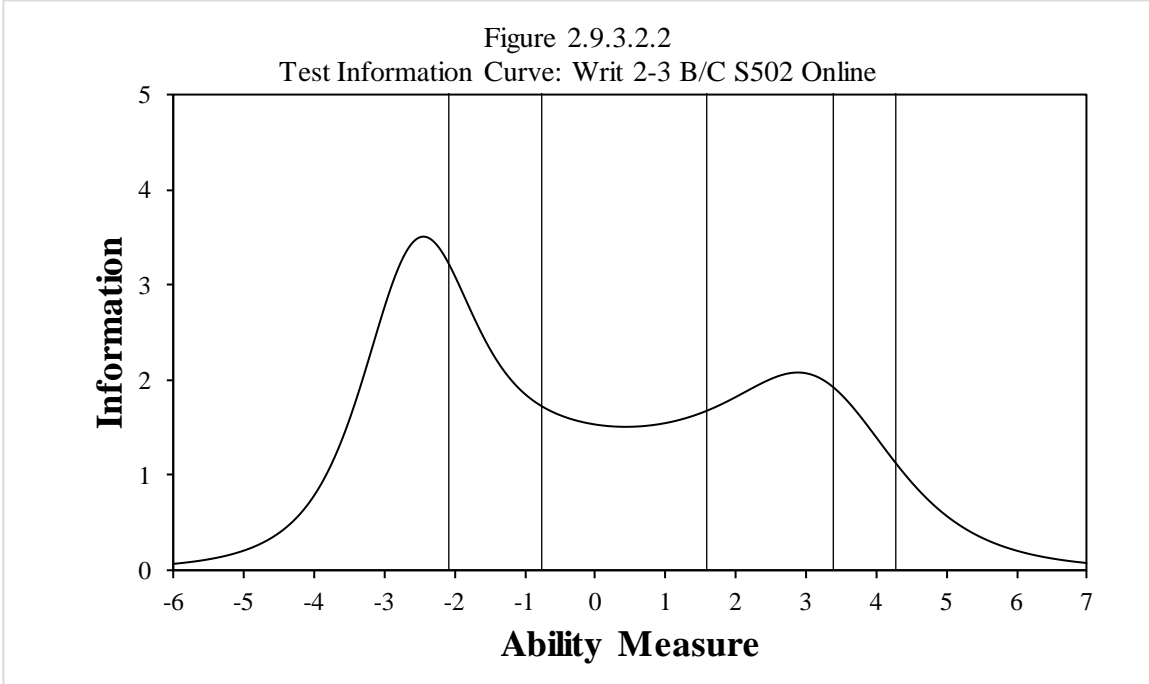
2.9.3.1 Grade 1



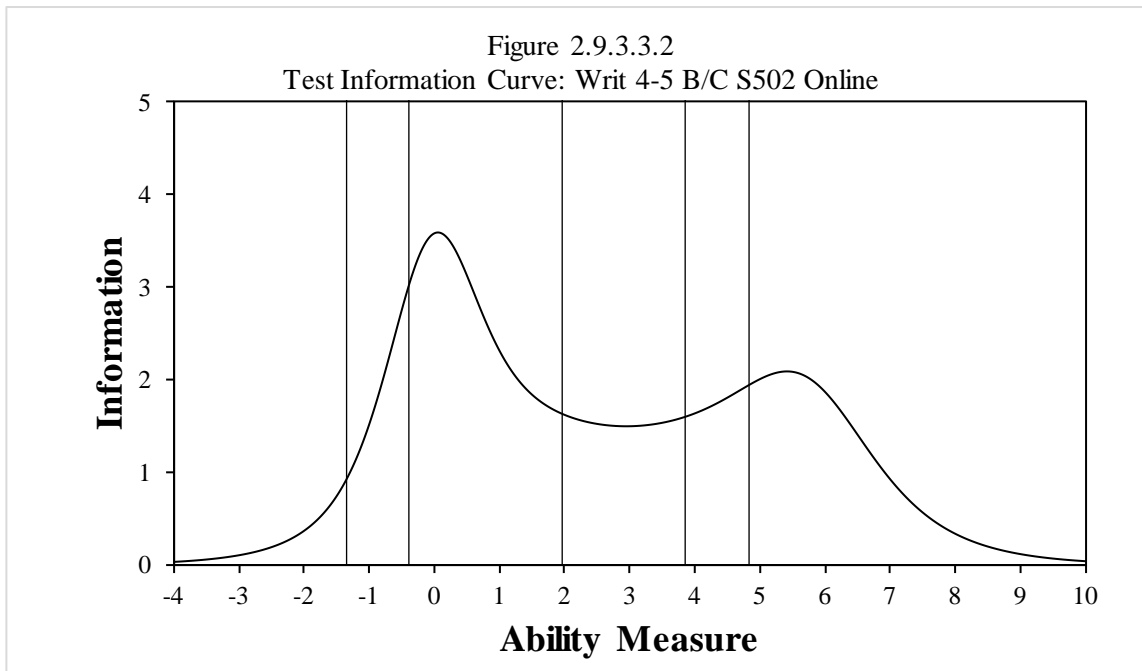
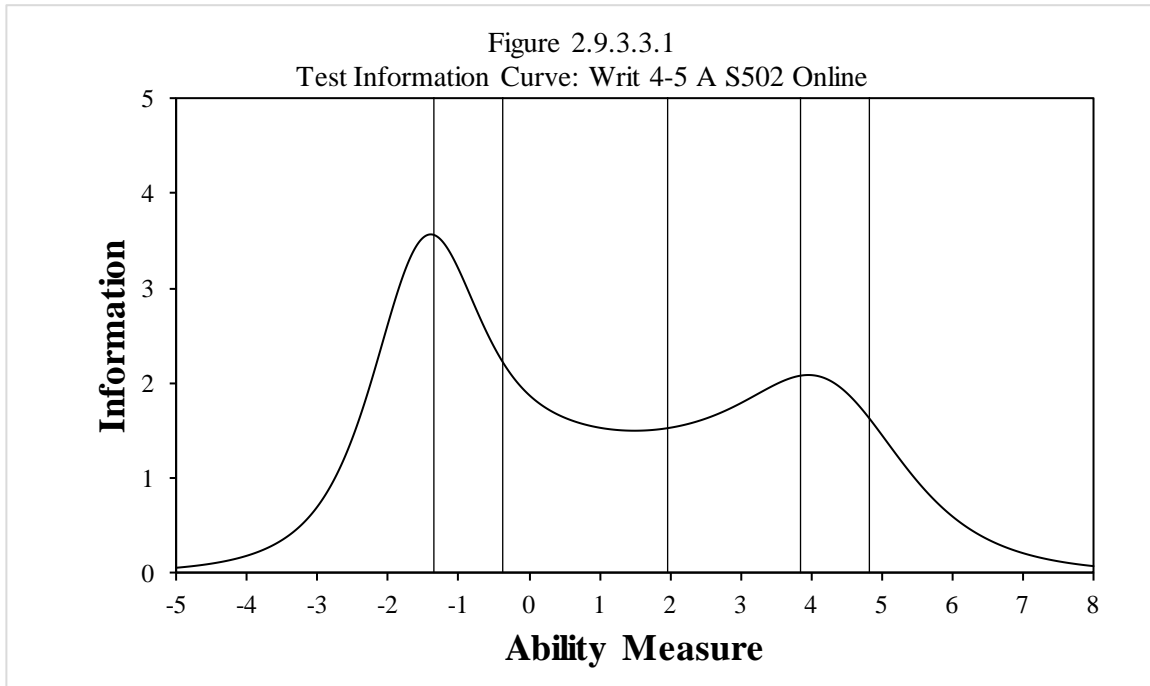


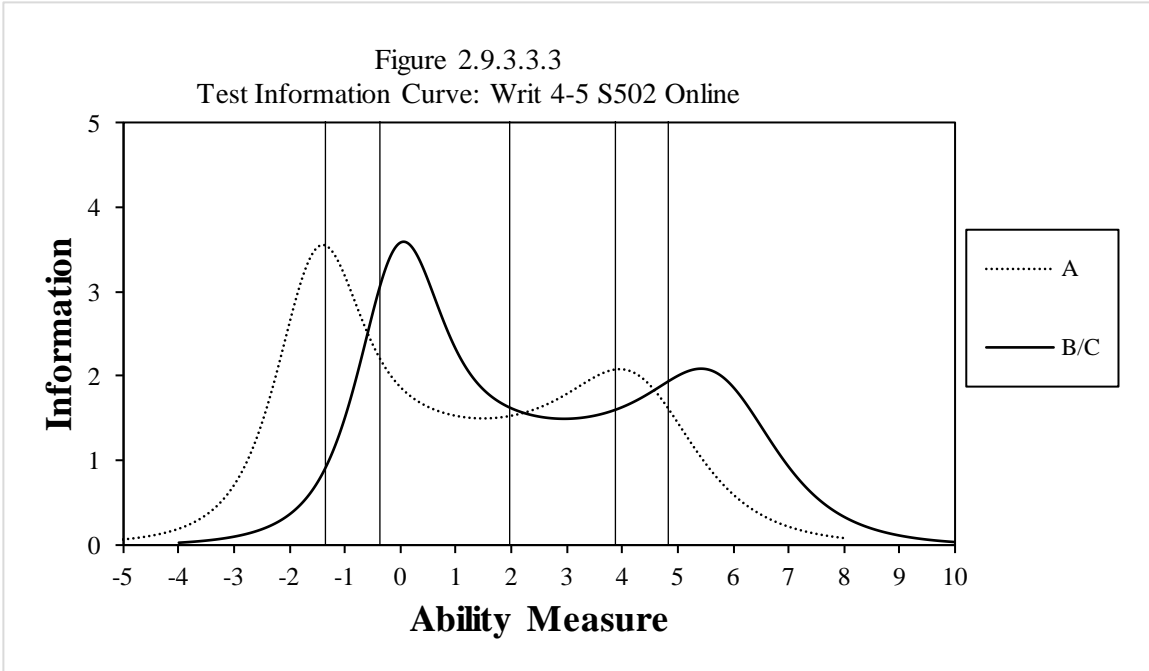
2.9.3.2 Grades 2–3



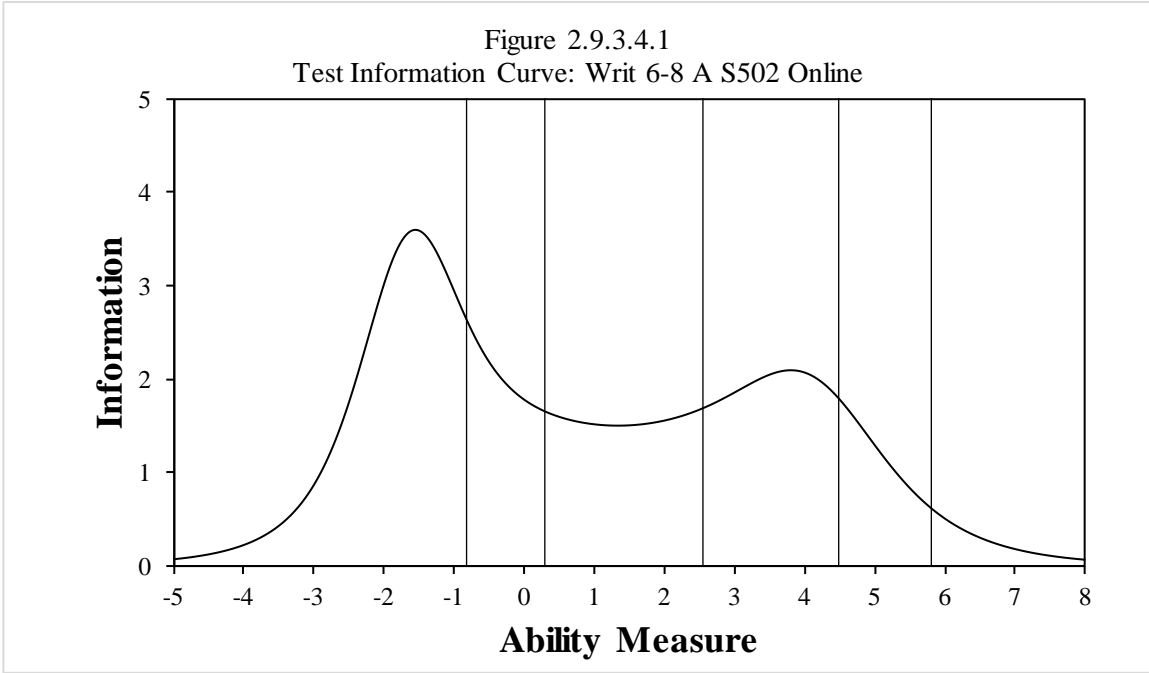


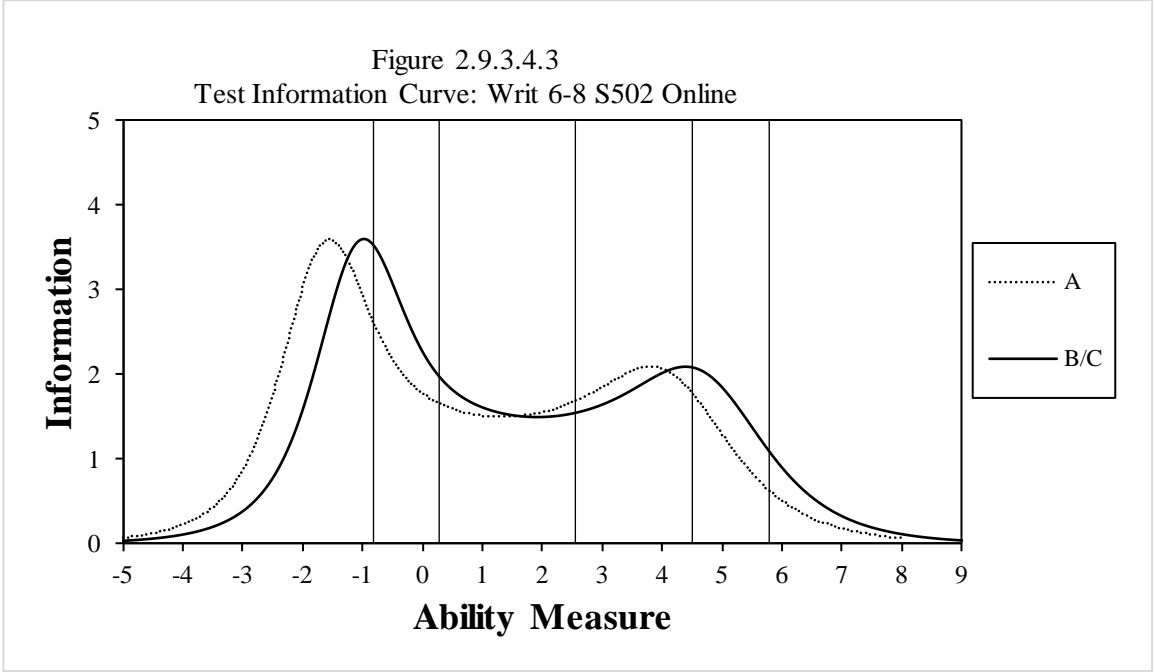
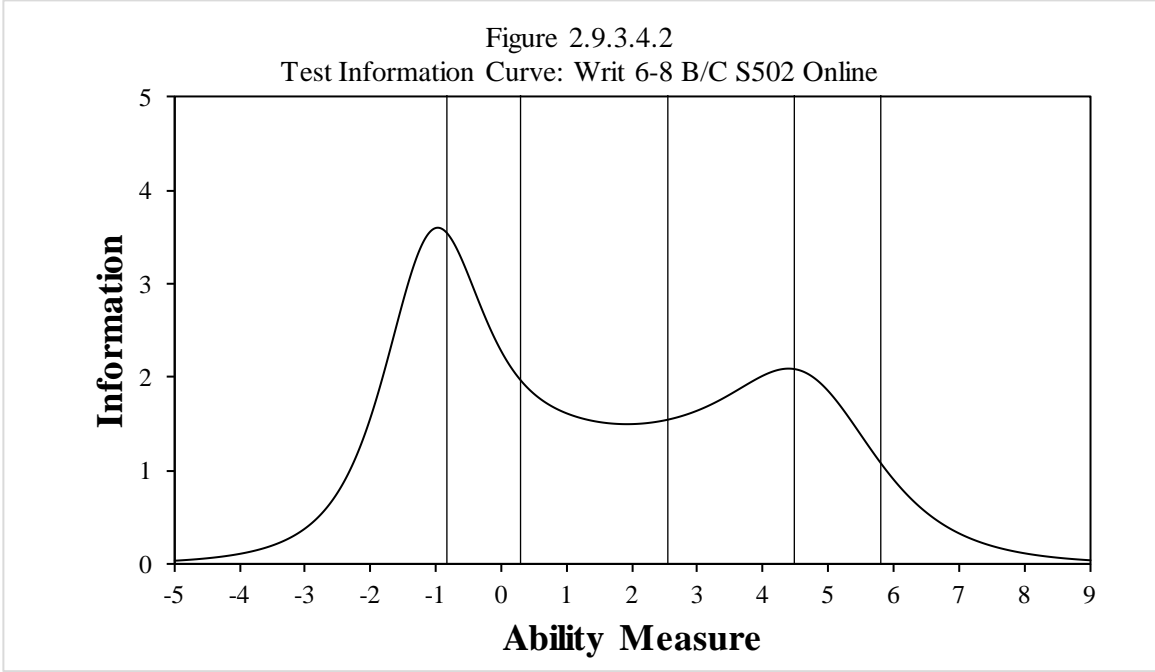
2.9.3.3 Grades 4–5



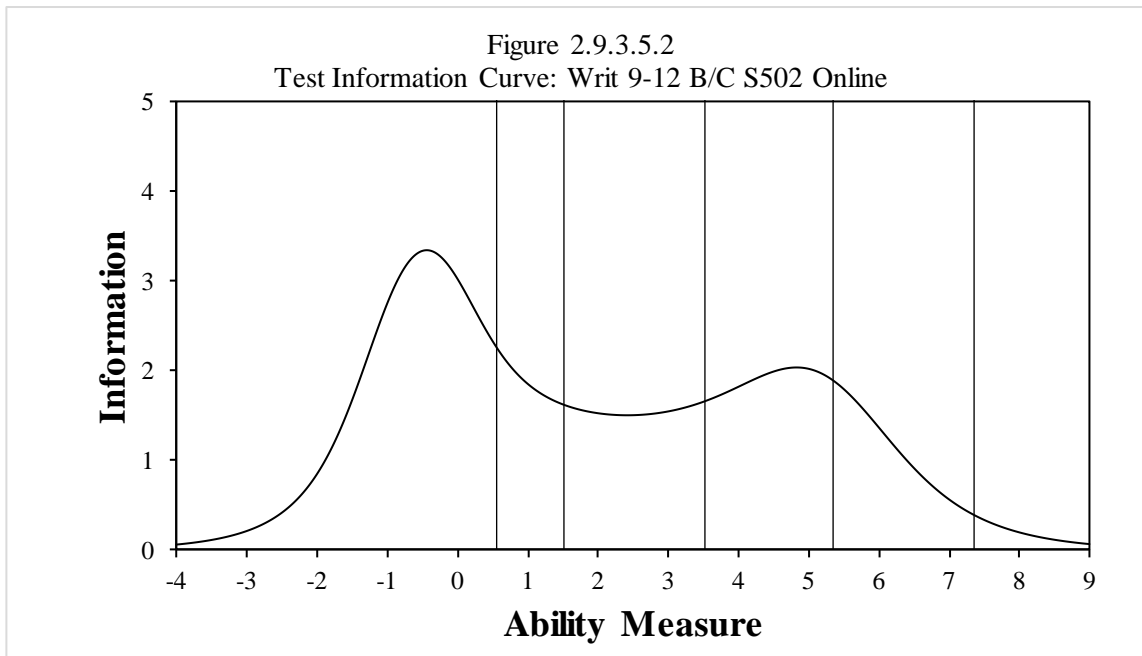
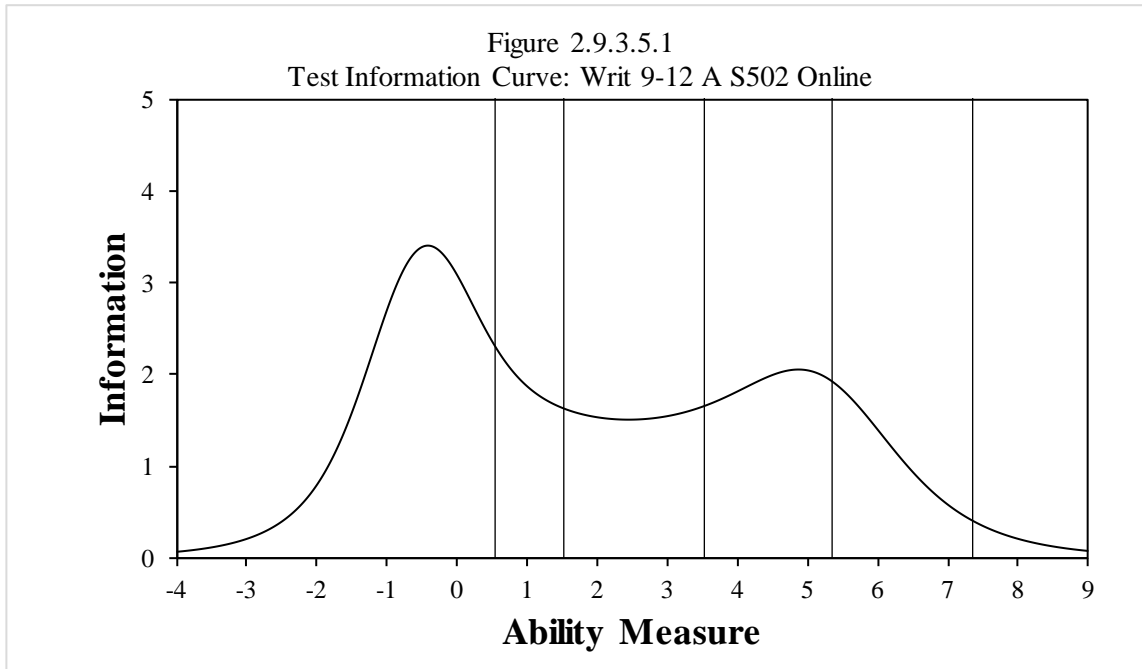


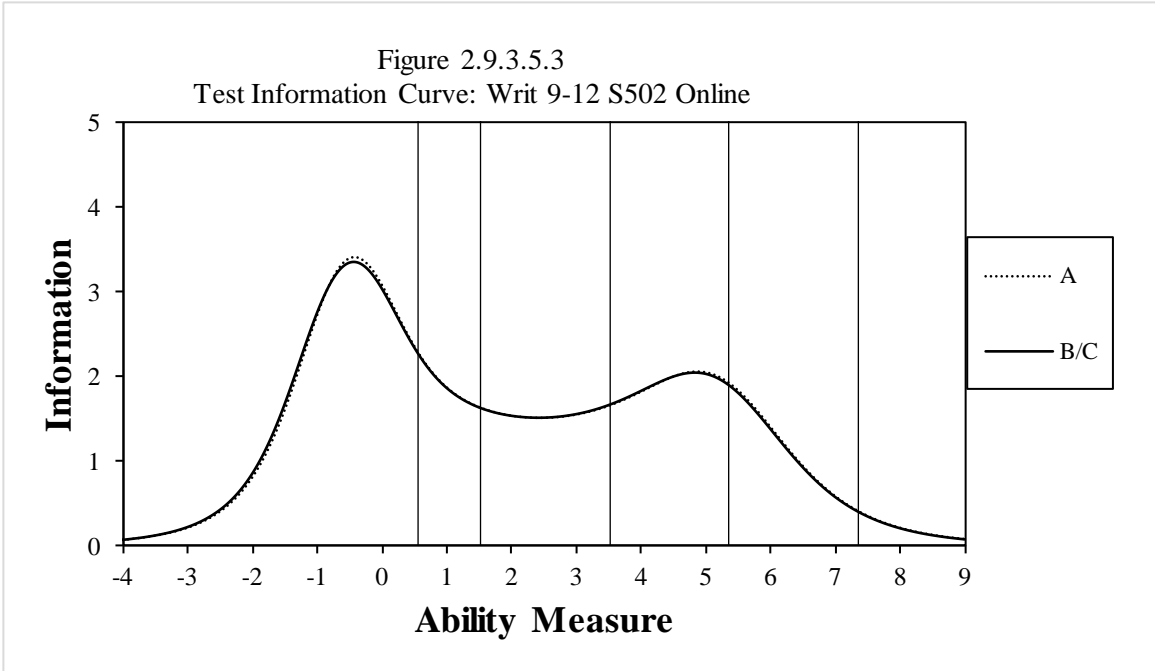
2.9.3.4 Grades 6–8





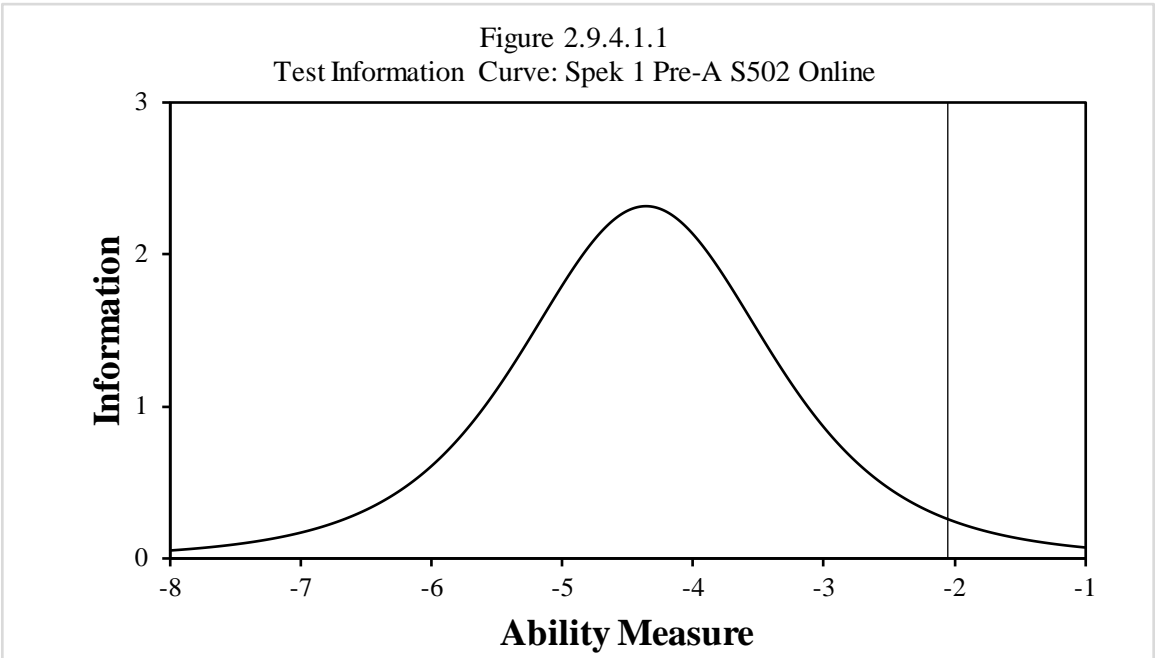
2.9.3.5 Grades 9-12

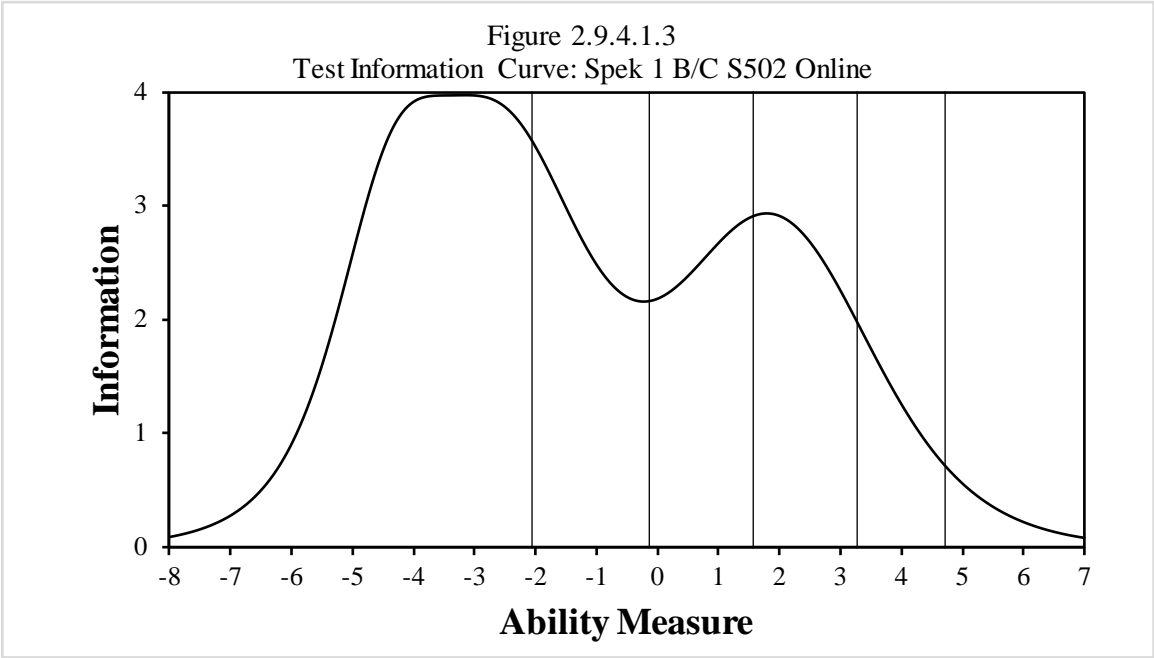
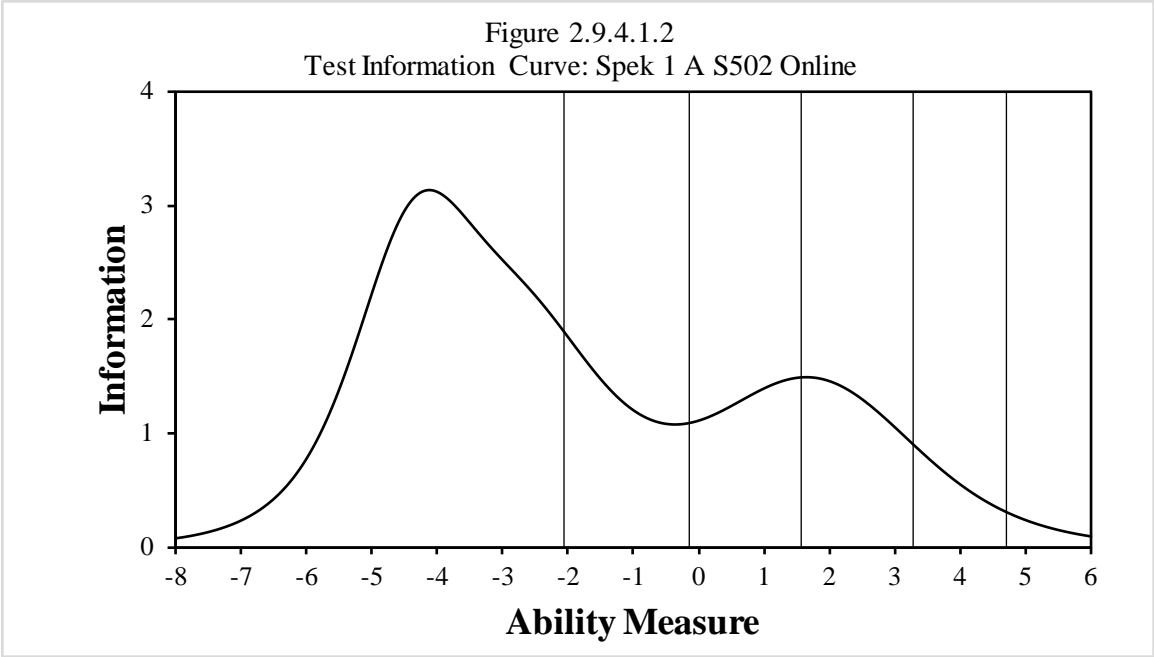


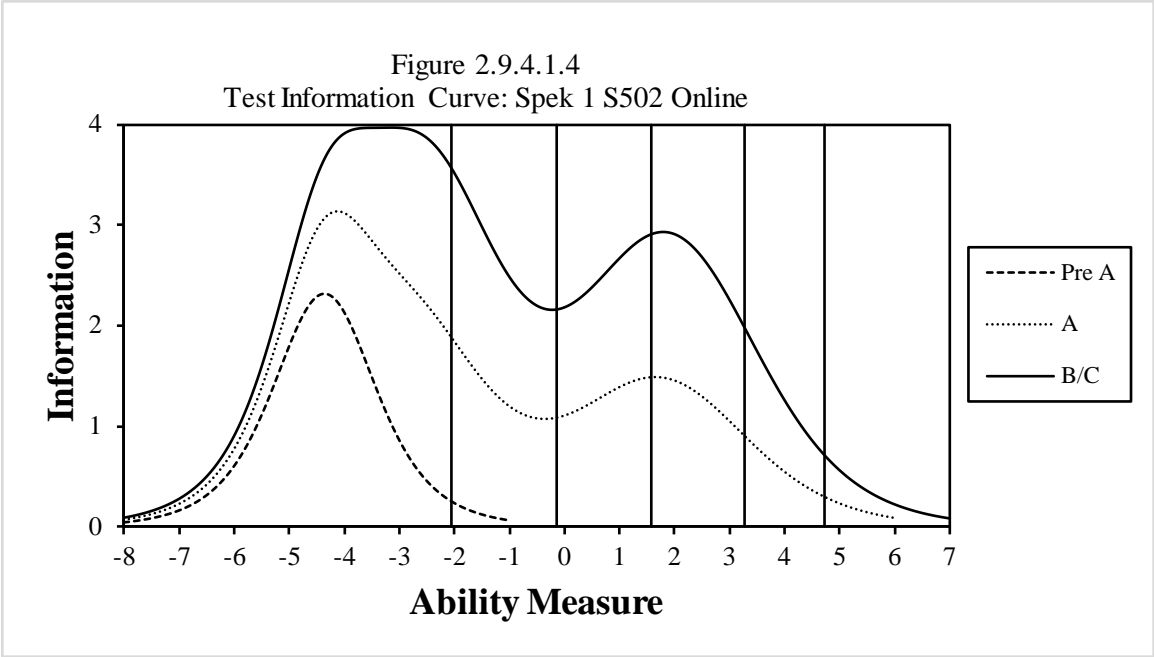


2.9.4 Speaking

2.9.4.1 Grade 1







2.9.4.2 Grades 2–3

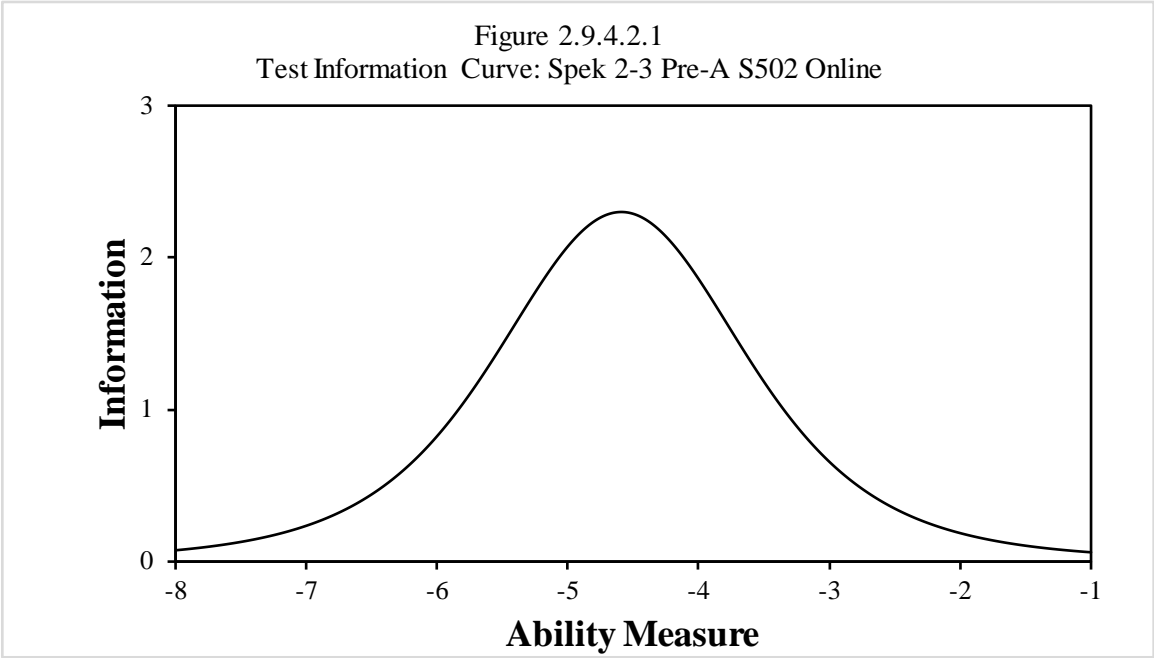


Figure 2.9.4.2.2
Test Information Curve: Spek 2-3 A S502 Online

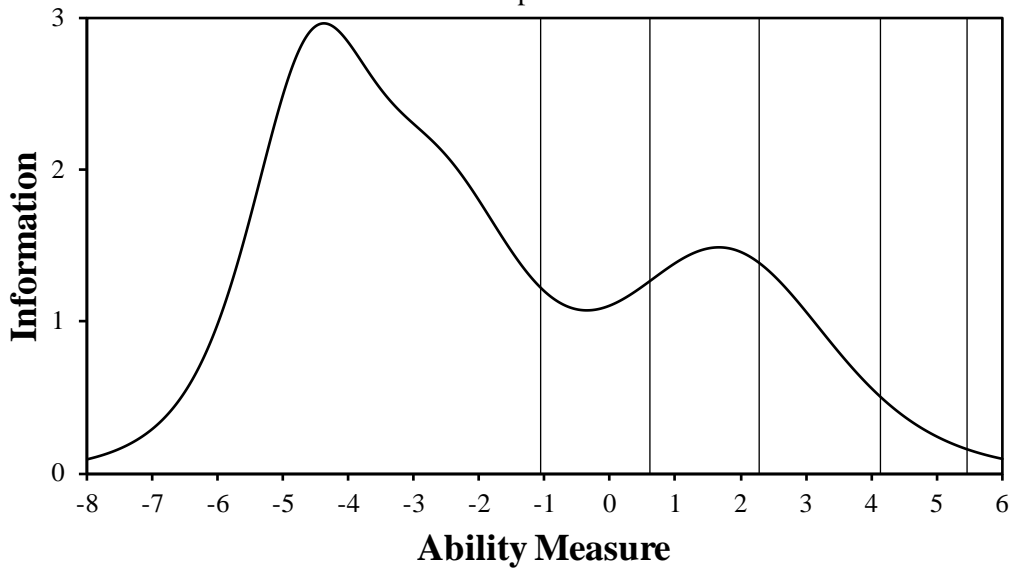
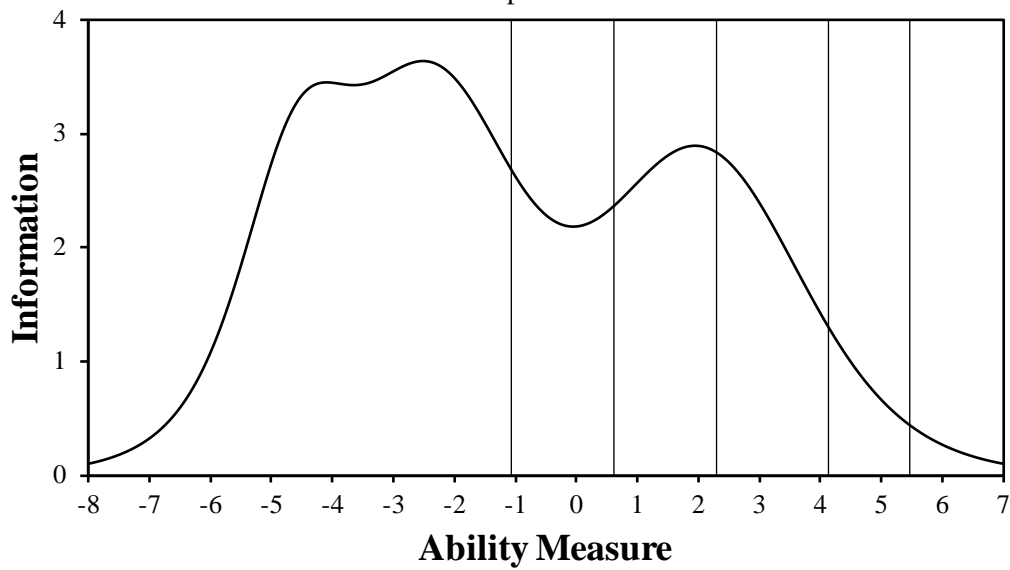
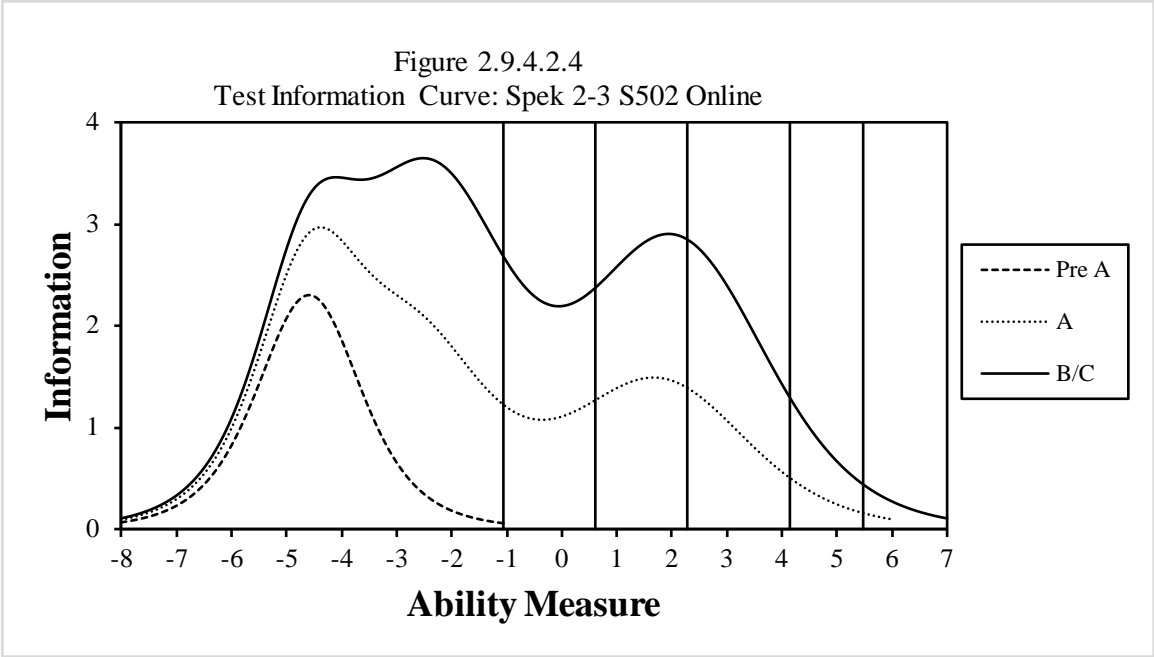
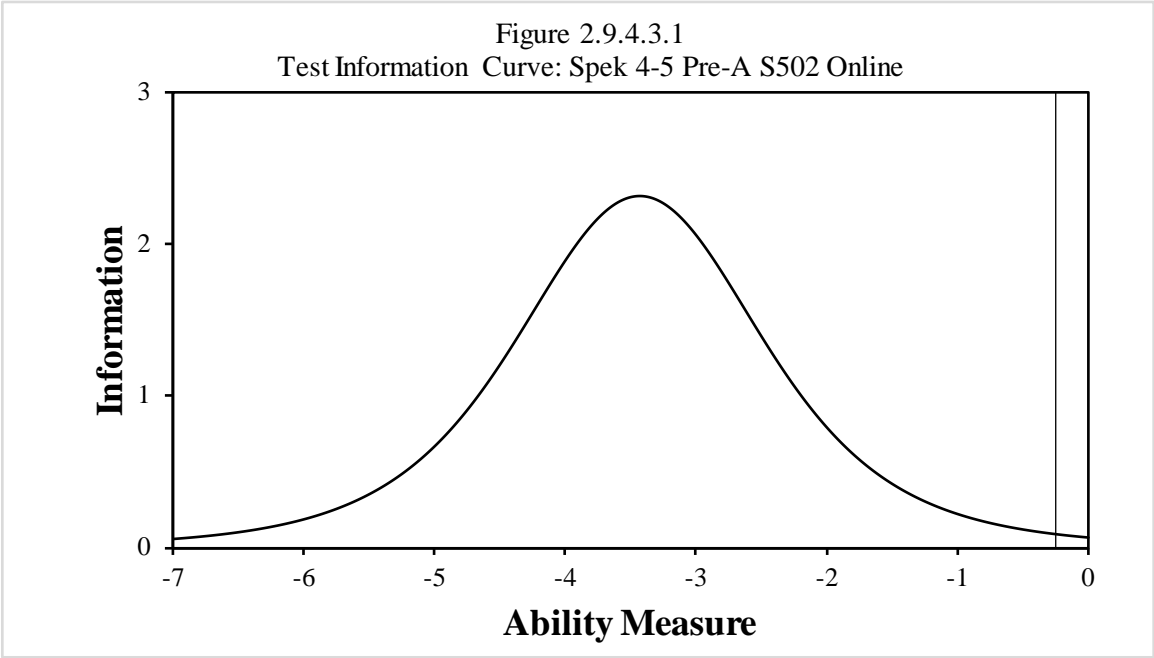


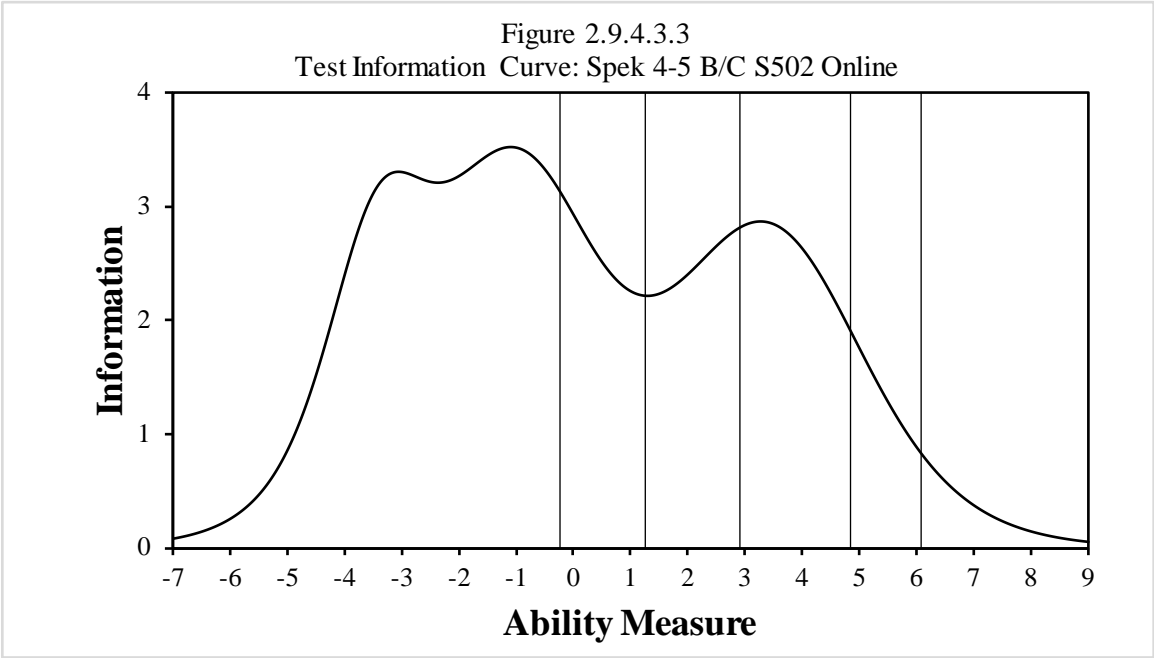
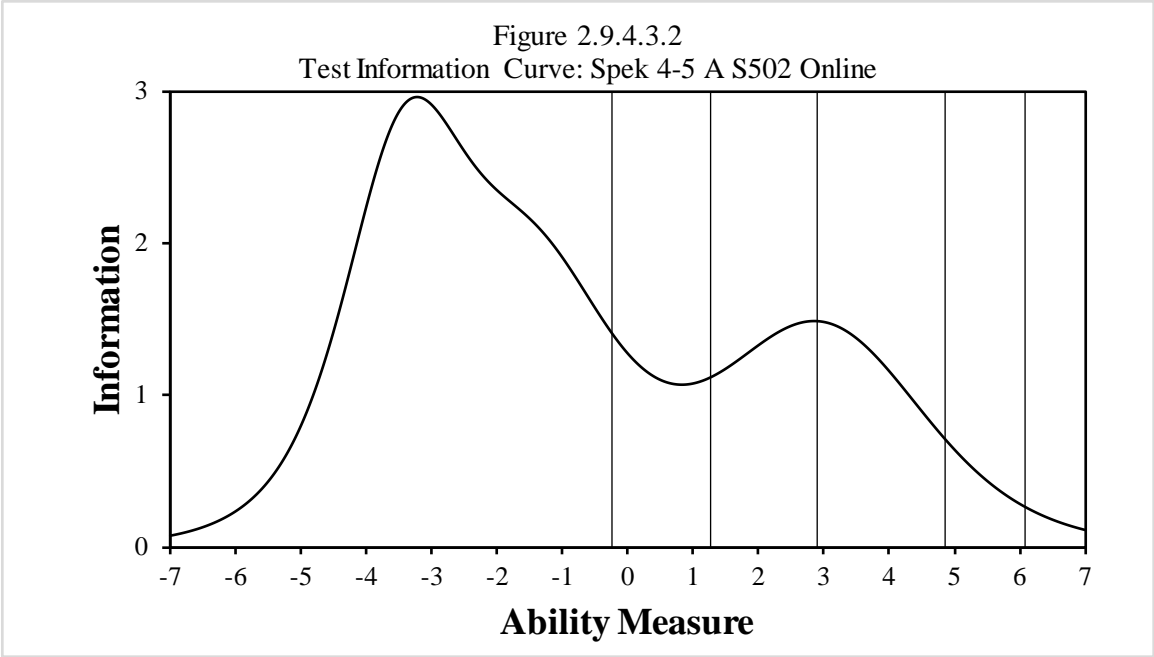
Figure 2.9.4.2.3
Test Information Curve: Spek 2-3 B/C S502 Online

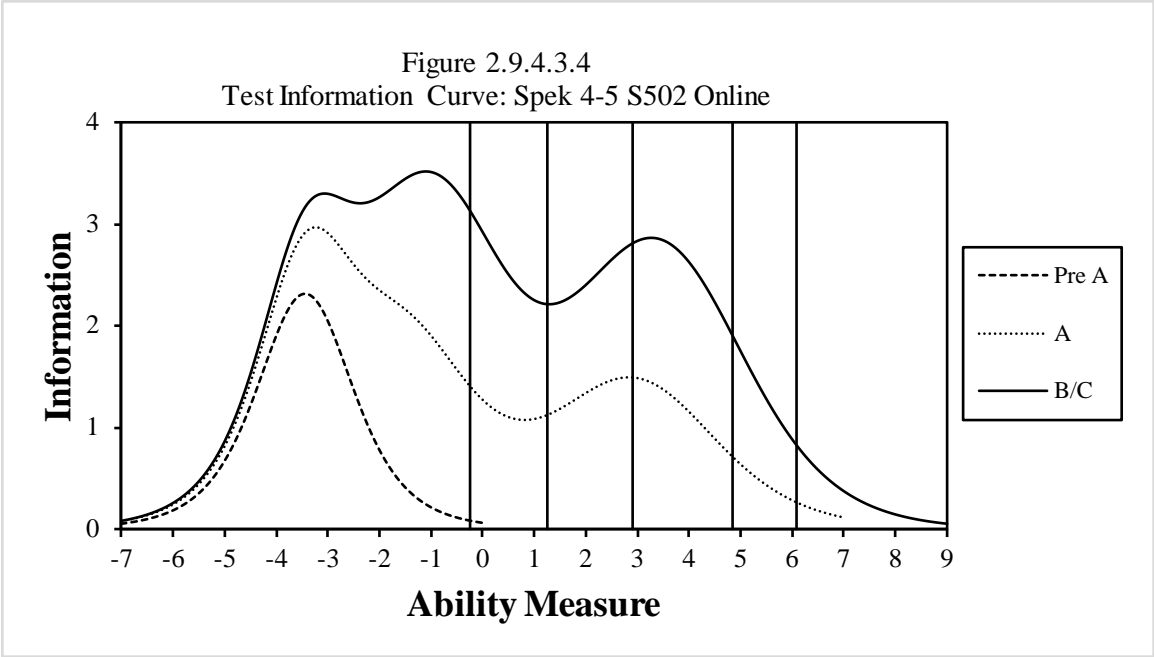




2.9.4.3 Grades 4–5







2.9.4.4 Grades 6–8

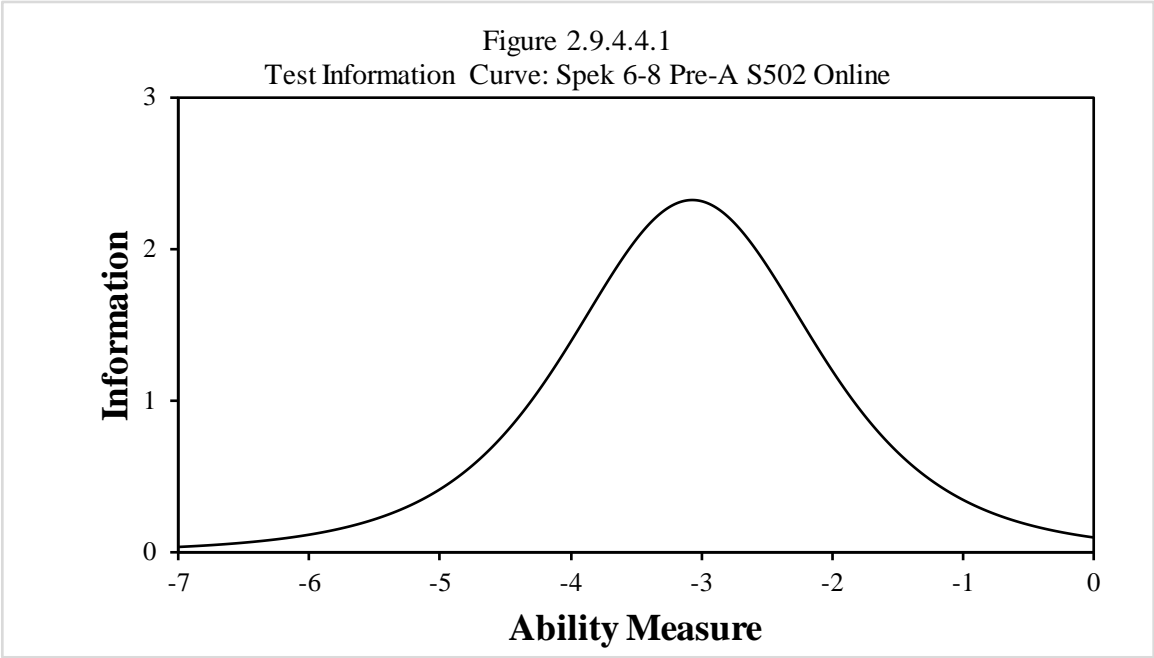


Figure 2.9.4.4.2
Test Information Curve: Spek 6-8 A S502 Online

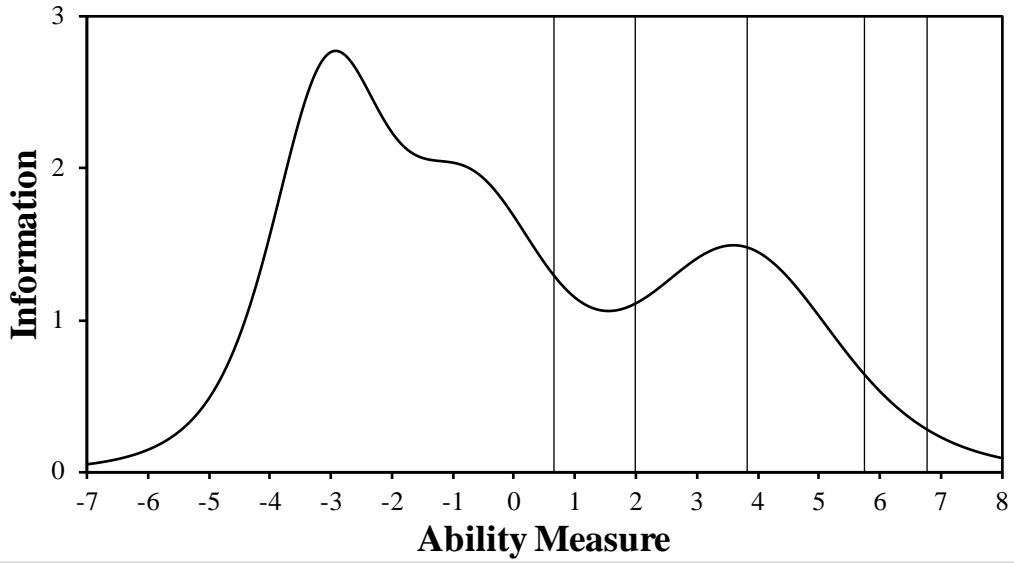
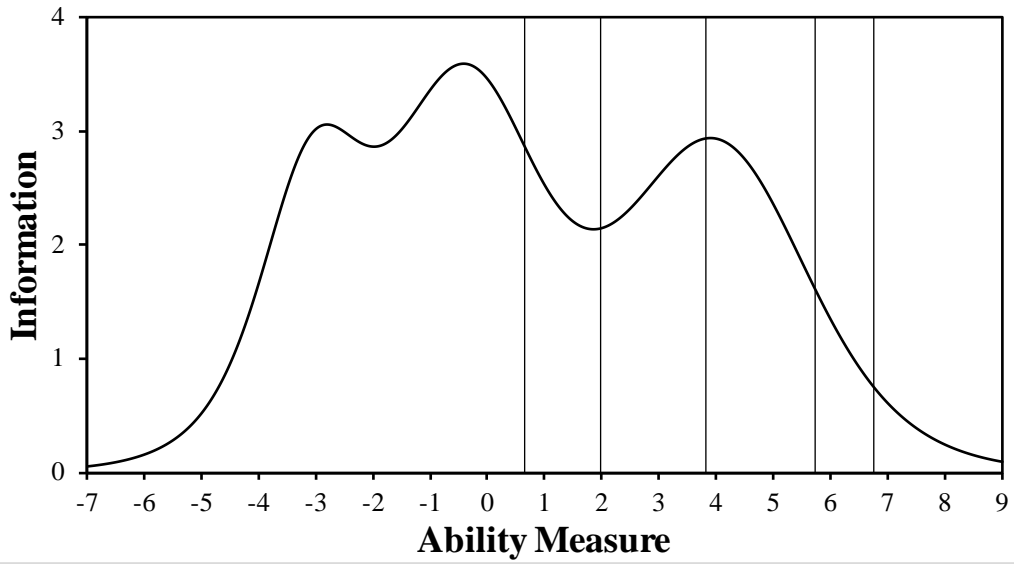
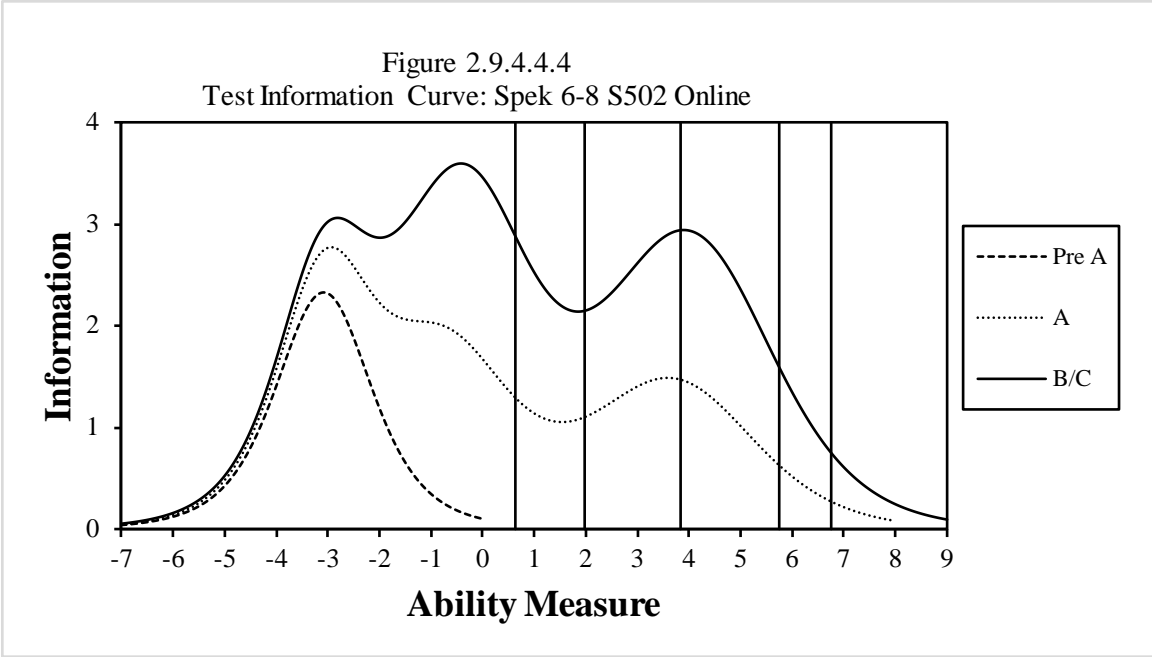
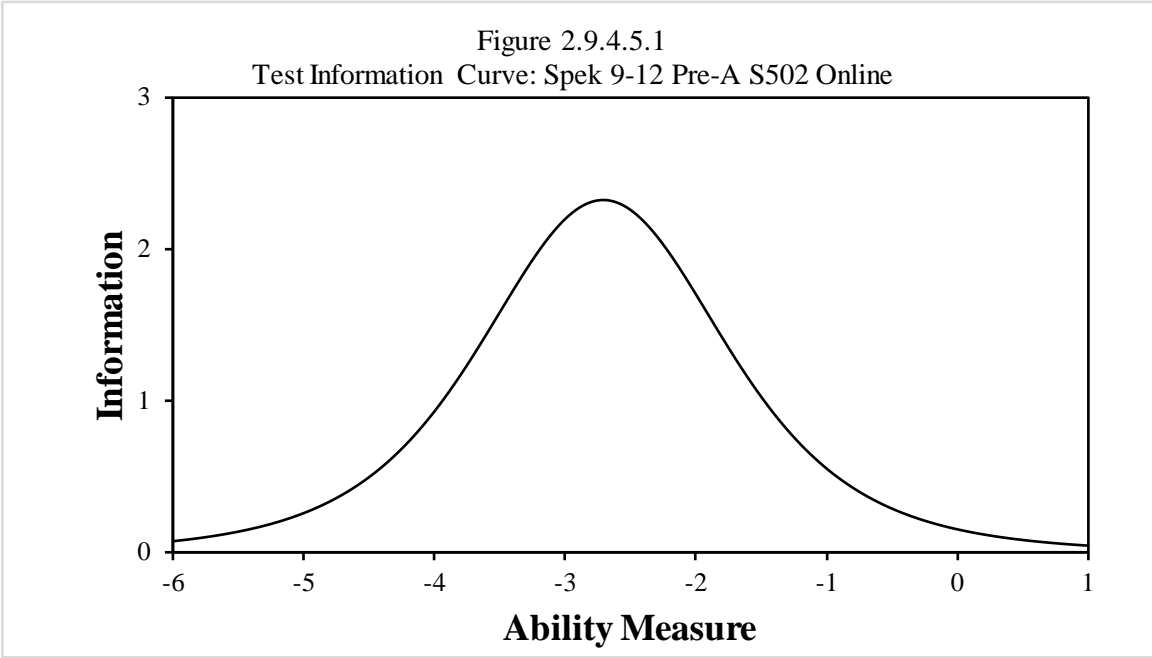


Figure 2.9.4.4.3
Test Information Curve: Spek 6-8 B/C S502 Online





2.9.4.5 Grades 9-12



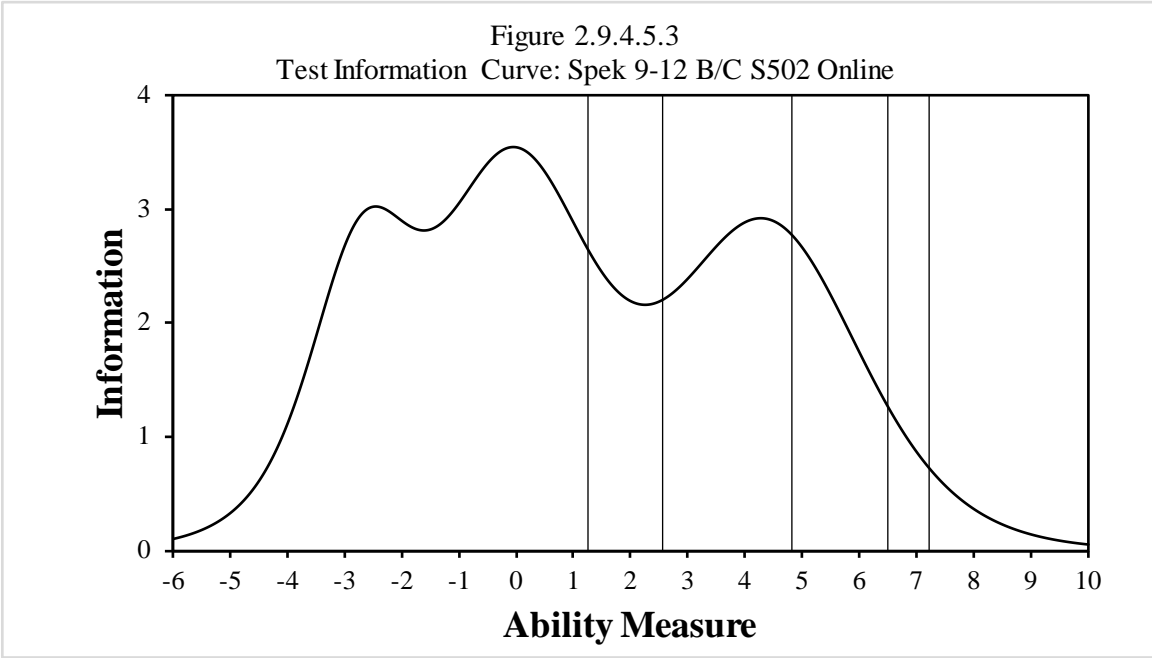
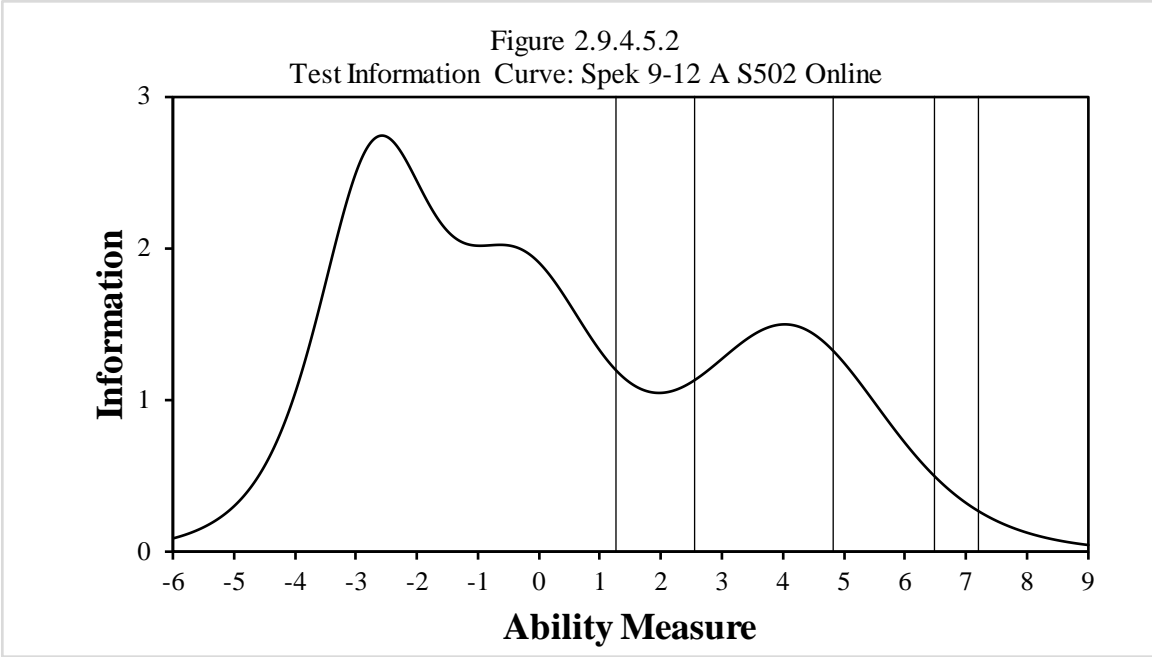
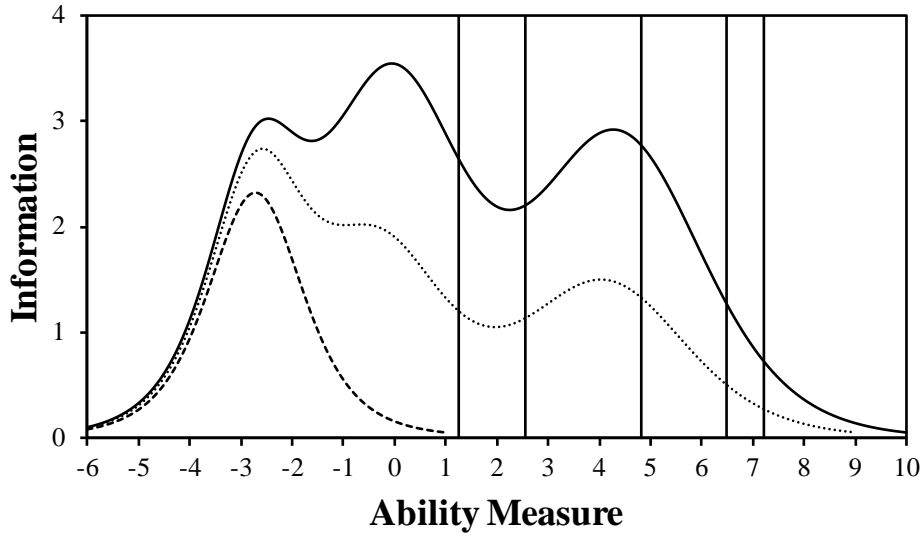


Figure 2.9.4.5.4
Test Information Curve: Spek 9-12 S502 Online



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

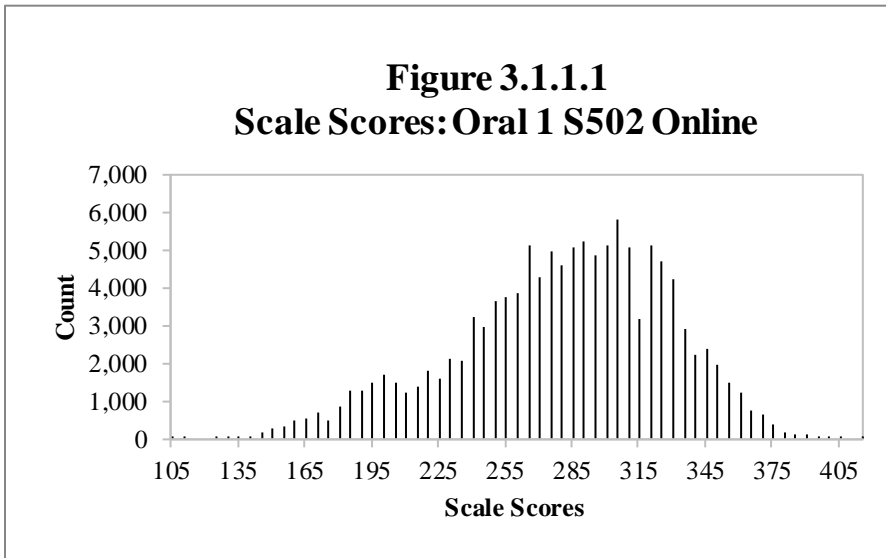
3.1.1 Oral

3.1.1.1 Grade 1

Table 3.1.1.1

Scale Score Descriptive Statistics: Oral 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	120,783	105	416	281.97	47.33
Total	120,783	105	416	281.97	47.33

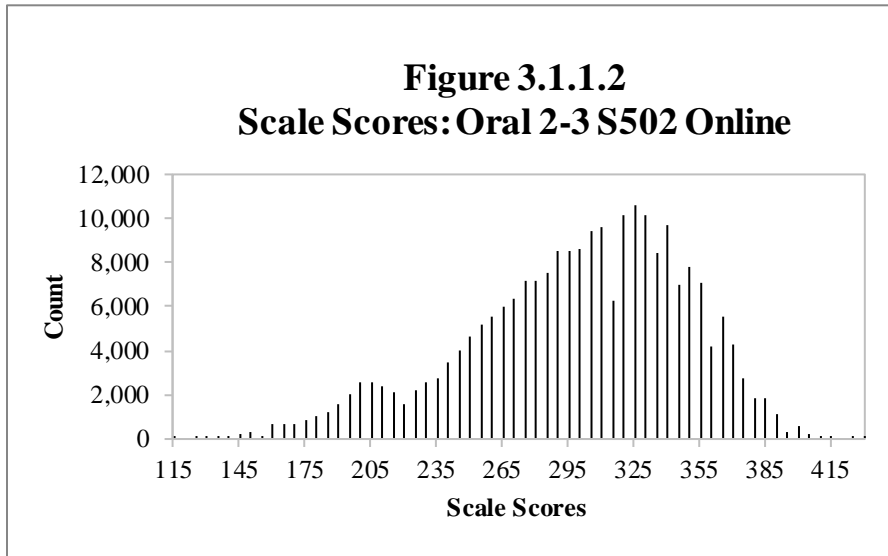


3.1.1.2 Grades 2–3

Table 3.1.1.2

Scale Score Descriptive Statistics: Oral 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	114,298	115	434	290.35	47.81
3	115,203	115	434	310.80	49.96
Total	229,501	115	434	300.61	49.96

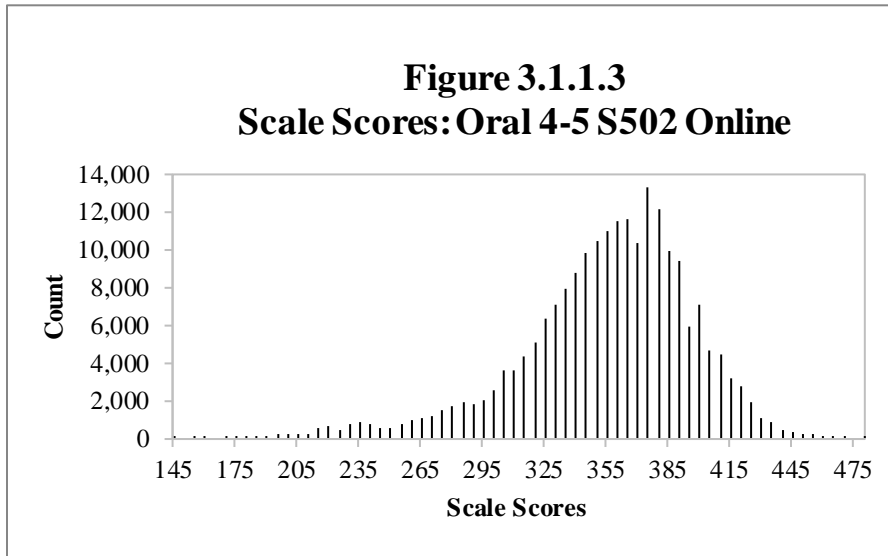


3.1.1.3 Grades 4–5

Table 3.1.1.3

Scale Score Descriptive Statistics: Oral 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	118,071	147	481	355.69	41.75
5	93,982	147	481	357.39	43.97
Total	212,053	147	481	356.44	42.76



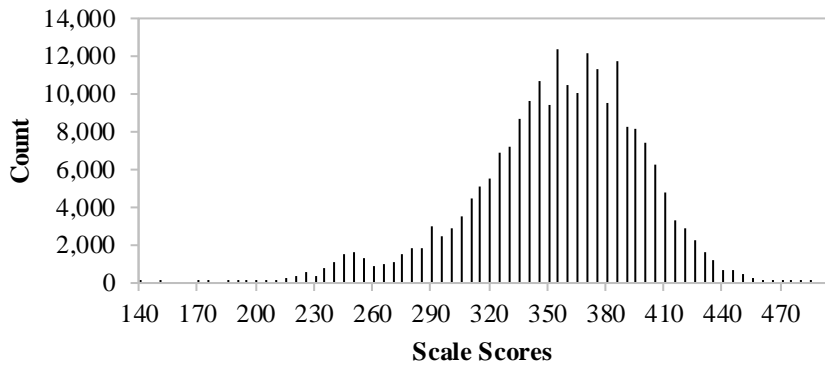
3.1.1.4 Grades 6–8

Table 3.1.1.4

Scale Score Descriptive Statistics: Oral 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	76,048	173	485	352.46	38.87
7	76,214	140	489	357.49	42.76
8	68,857	154	496	361.06	46.39
Total	221,119	140	496	356.87	42.80

Figure 3.1.1.4
Scale Scores: Oral 6-8 S502 Online

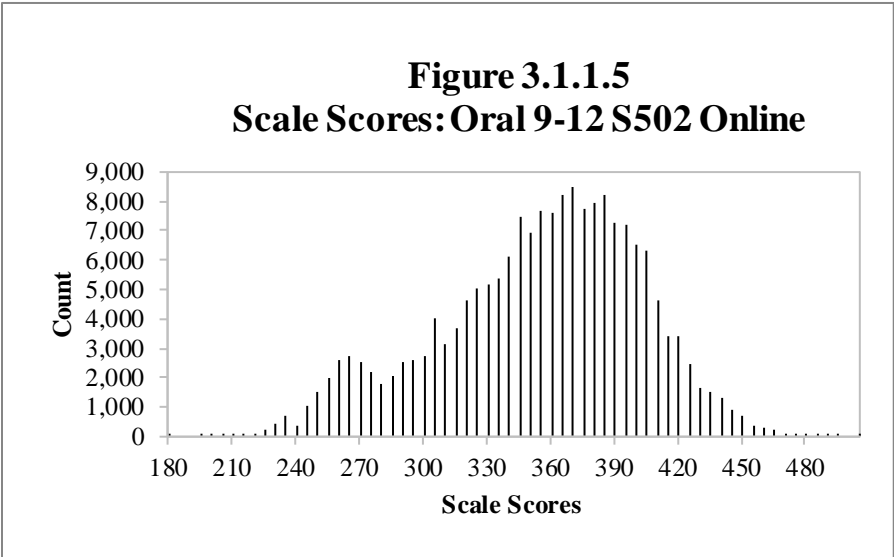


3.1.1.5 Grades 9-12

Table 3.1.1.5

Scale Score Descriptive Statistics: Oral 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	57,868	182	498	352.89	46.37
10	51,854	195	506	353.86	48.81
11	42,004	208	506	360.84	47.43
12	32,401	202	506	362.17	47.37
Total	184,127	182	506	356.61	47.65



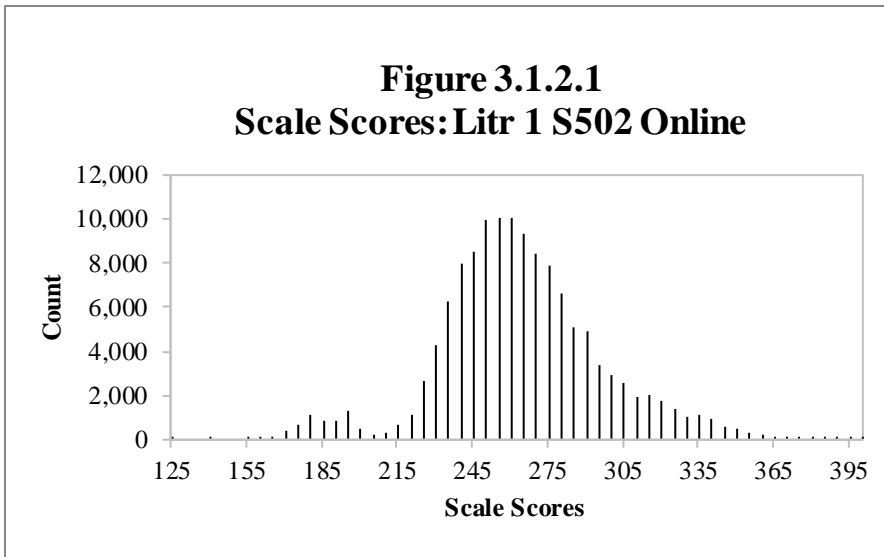
3.1.2 Literacy

3.1.2.1 Grade 1

Table 3.1.2.1

Scale Score Descriptive Statistics: Litr 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	131,049	126	400	265.79	32.42
Total	131,049	126	400	265.79	32.42

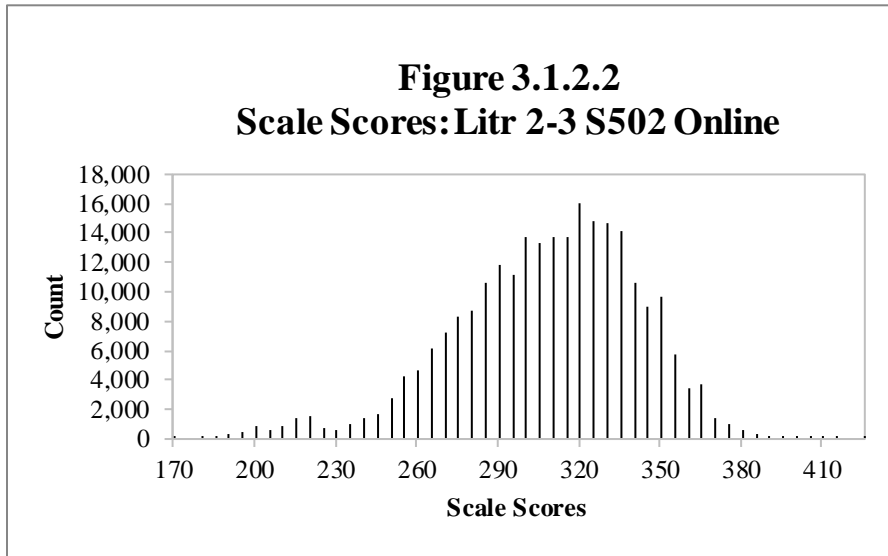


3.1.2.2 Grades 2–3

Table 3.1.2.2

Scale Score Descriptive Statistics: Litr 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	123,791	172	407	300.95	32.47
3	123,063	172	427	317.61	33.50
Total	246,854	172	427	309.25	34.02

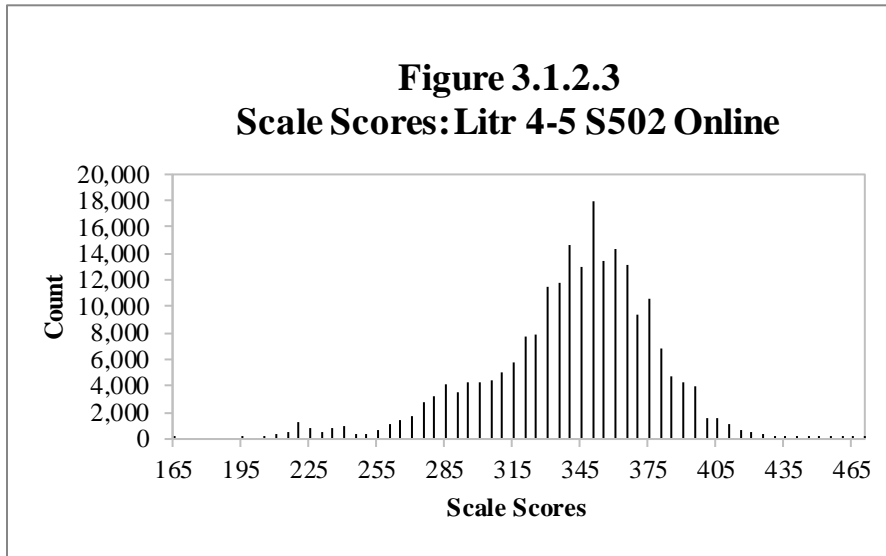


3.1.2.3 Grades 4–5

Table 3.1.2.3

Scale Score Descriptive Statistics: Litr 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	122,315	165	458	338.84	36.18
5	96,943	165	471	345.44	37.17
Total	219,258	165	471	341.76	36.77

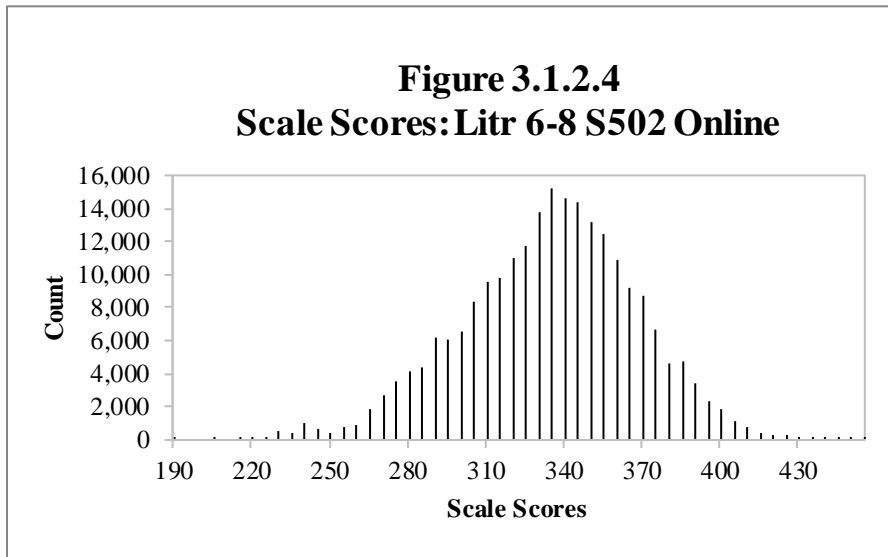


3.1.2.4 Grades 6–8

Table 3.1.2.4

Scale Score Descriptive Statistics: Litr 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	79,491	209	450	328.17	29.85
7	78,845	209	459	337.46	33.17
8	70,630	194	455	342.89	35.68
Total	228,966	194	459	335.91	33.43



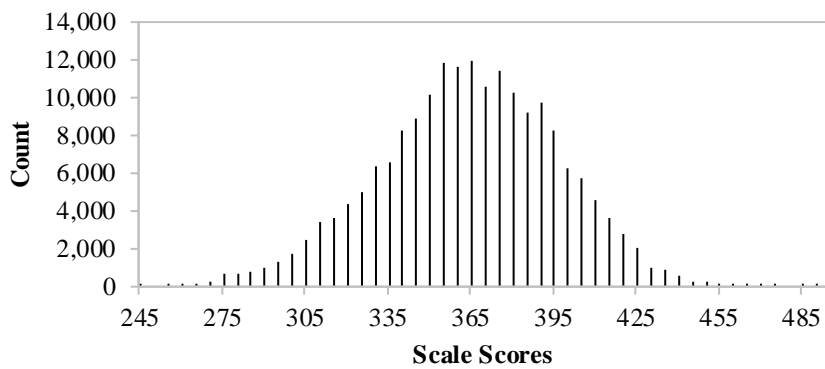
3.1.2.5 Grades 9-12

Table 3.1.2.5

Scale Score Descriptive Statistics: Litr 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	59,411	247	471	361.39	32.44
10	53,046	247	476	363.77	33.19
11	43,143	247	496	369.98	31.81
12	33,142	257	492	372.13	30.81
Total	188,742	247	496	365.91	32.51

Figure 3.1.2.5
Scale Scores: Litr 9-12 S502 Online



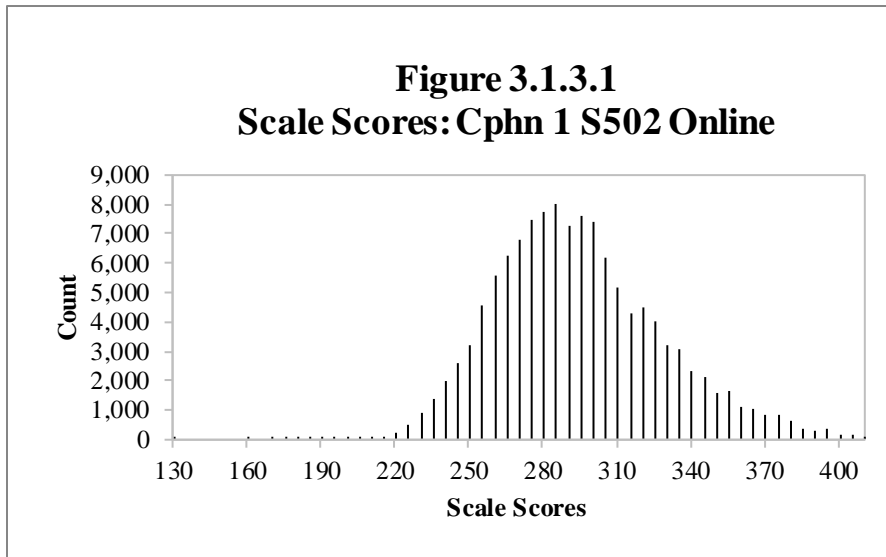
3.1.3 Comprehension

3.1.3.1 Grade 1

Table 3.1.3.1

Scale Score Descriptive Statistics: Cphn 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	123,489	130	414	295.97	33.66
Total	123,489	130	414	295.97	33.66

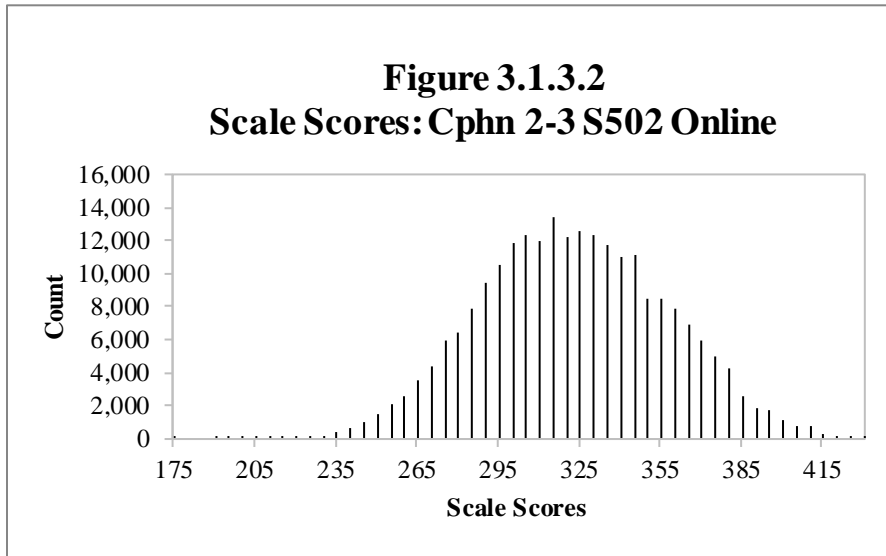


3.1.3.2 Grades 2–3

Table 3.1.3.2

Scale Score Descriptive Statistics: Cphn 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	116,639	194	432	317.54	31.33
3	116,528	175	432	331.73	36.01
Total	233,167	175	432	324.63	34.49

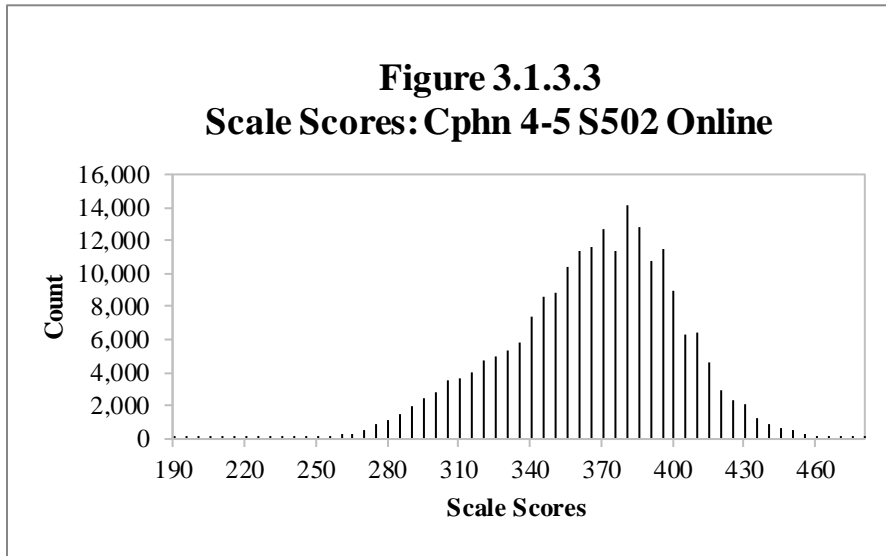


3.1.3.3 Grades 4–5

Table 3.1.3.3

Scale Score Descriptive Statistics: Cphn 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	118,624	196	480	365.61	34.67
5	94,298	192	480	369.90	36.61
Total	212,922	192	480	367.51	35.61

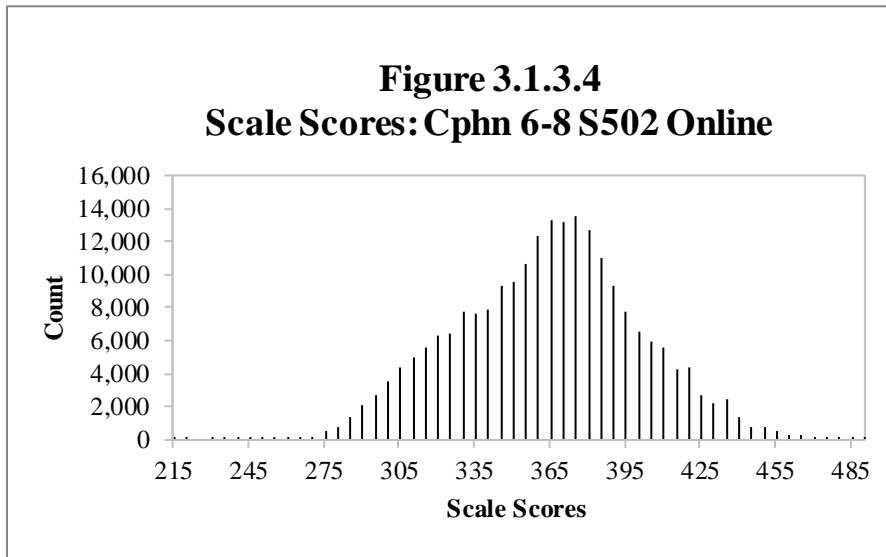


3.1.3.4 Grades 6–8

Table 3.1.3.4

Scale Score Descriptive Statistics: Cphn 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	76,997	223	492	358.73	32.02
7	76,867	235	492	367.00	36.34
8	69,310	217	492	372.15	39.76
Total	223,174	217	492	365.75	36.47

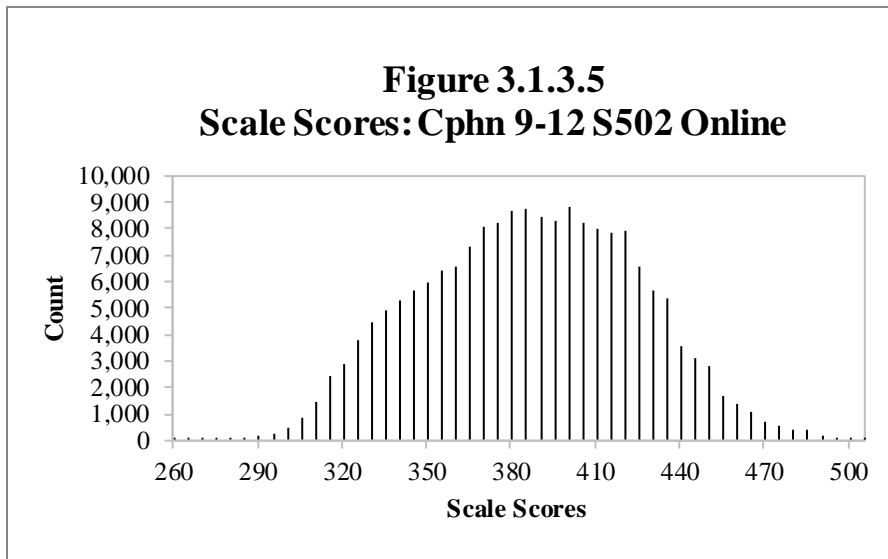


3.1.3.5 Grades 9-12

Table 3.1.3.5

Scale Score Descriptive Statistics: Cphn 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	57,710	263	506	383.96	36.99
10	51,830	271	506	386.43	38.96
11	42,228	267	506	392.40	38.42
12	32,427	271	506	394.38	37.96
Total	184,195	263	506	388.42	38.28



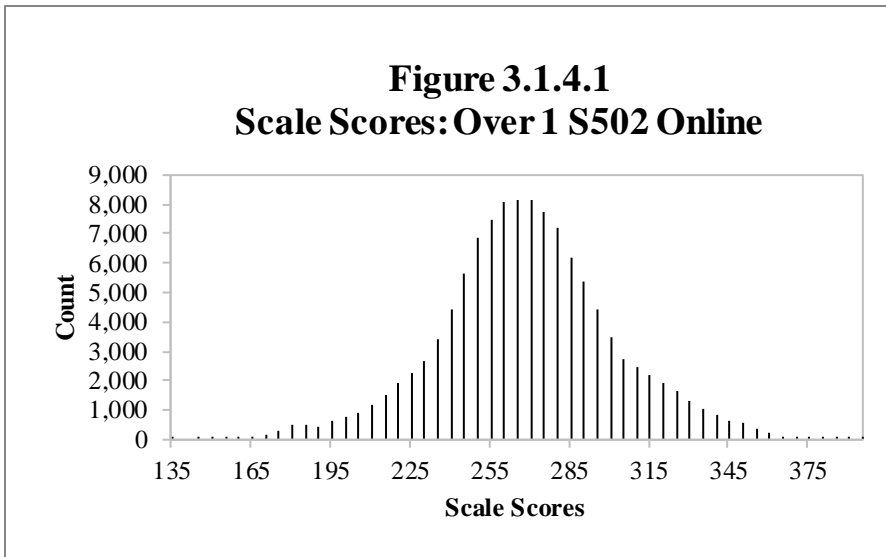
3.1.4 Overall

3.1.4.1 Grade 1

Table 3.1.4.1

Scale Score Descriptive Statistics: Over 1 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	116,586	135	397	270.55	32.75
Total	116,586	135	397	270.55	32.75

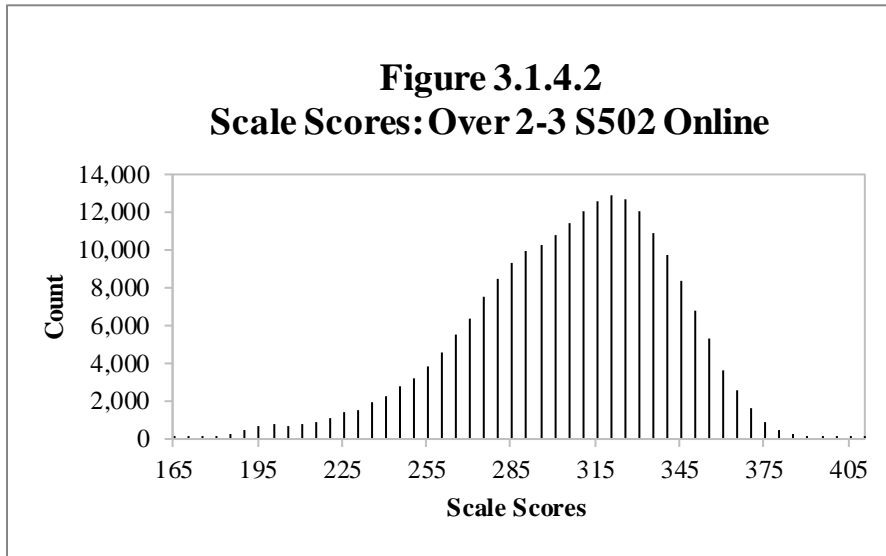


3.1.4.2 Grades 2–3

Table 3.1.4.2

Scale Score Descriptive Statistics: Over 2-3 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	109,241	168	401	297.65	33.87
3	110,093	165	412	315.40	35.63
Total	219,334	165	412	306.56	35.88

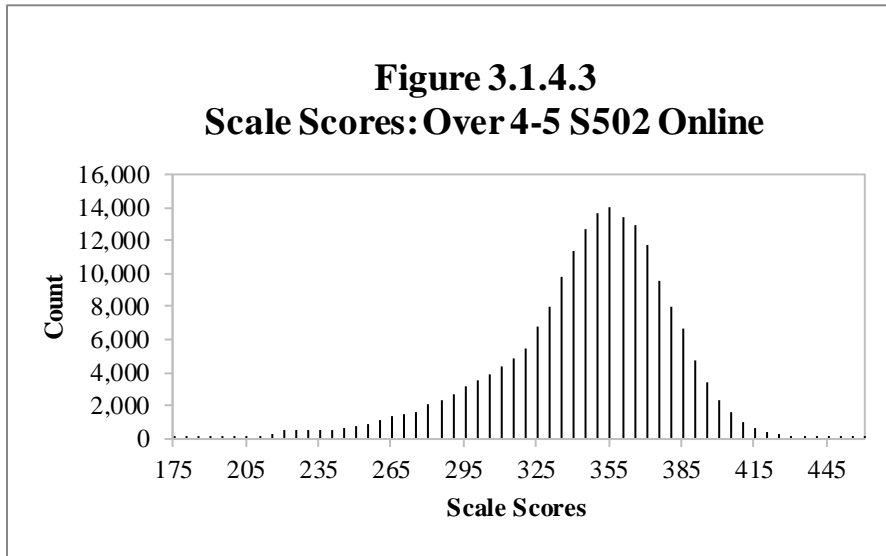


3.1.4.3 Grades 4–5

Table 3.1.4.3

Scale Score Descriptive Statistics: Over 4-5 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	109,168	175	455	343.86	35.29
5	87,209	184	461	348.95	36.63
Total	196,377	175	461	346.12	35.98

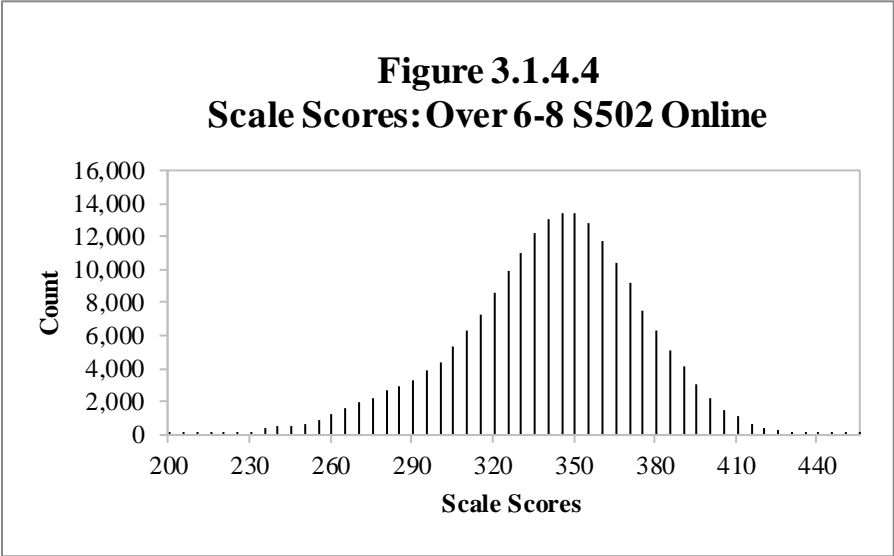


3.1.4.4 Grades 6–8

Table 3.1.4.4

Scale Score Descriptive Statistics: Over 6-8 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	70,490	217	450	335.38	30.00
7	70,293	206	457	343.43	33.73
8	63,624	200	458	348.08	36.66
Total	204,407	200	458	342.10	33.87

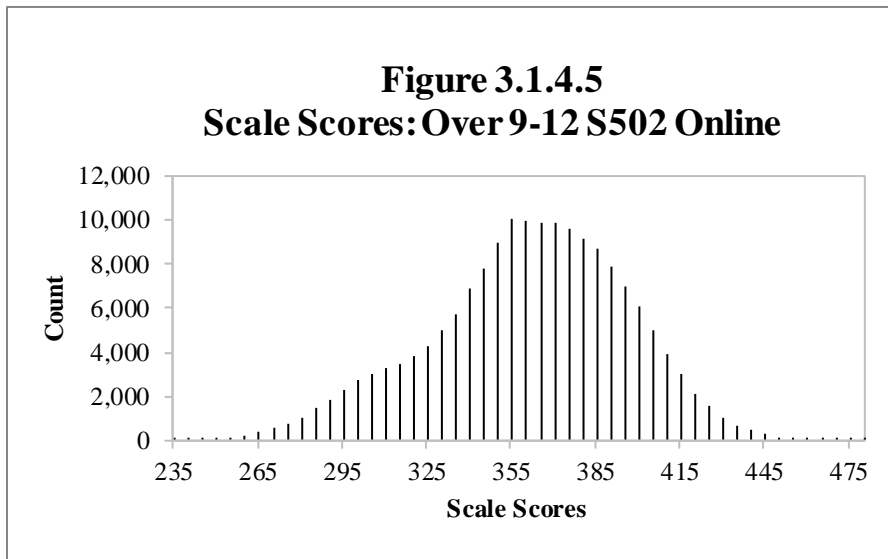


3.1.4.5 Grades 9-12

Table 3.1.4.5

Scale Score Descriptive Statistics: Over 9-12 S502 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	53,305	238	472	358.53	34.48
10	48,010	251	478	360.48	35.90
11	38,947	248	483	366.91	34.35
12	30,090	248	481	368.83	33.29
Total	170,352	238	483	362.81	34.90



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

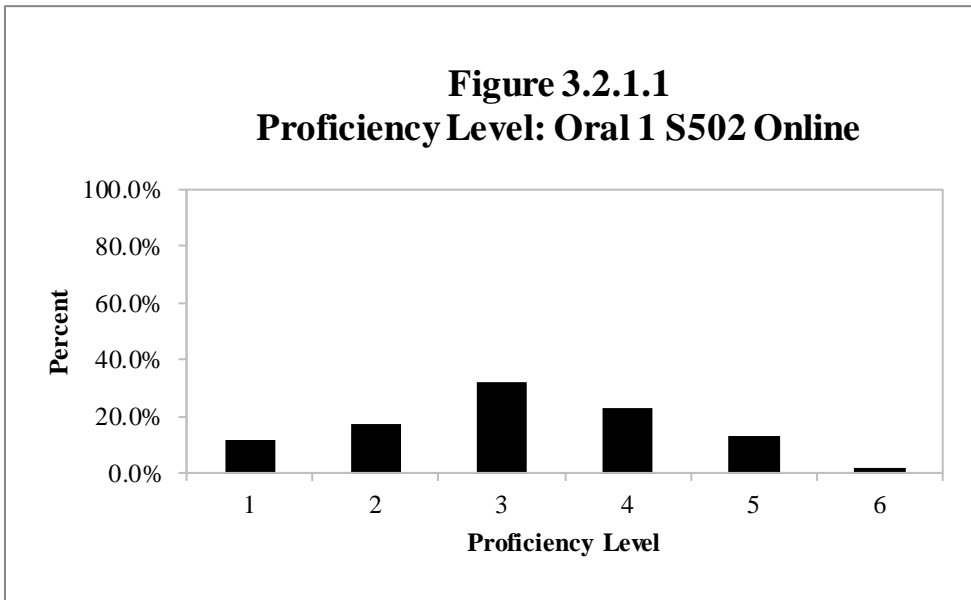
3.2.1 Oral

3.2.1.1 Grade 1

Table 3.2.1.1

Proficiency Level Distribution: Oral 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	14,073	11.65%	14,073	11.65%
2	20,921	17.32%	20,921	17.32%
3	39,274	32.52%	39,274	32.52%
4	27,990	23.17%	27,990	23.17%
5	16,225	13.43%	16,225	13.43%
6	2,300	1.90%	2,300	1.90%
Total	120,783	100.00%	120,783	100.00%

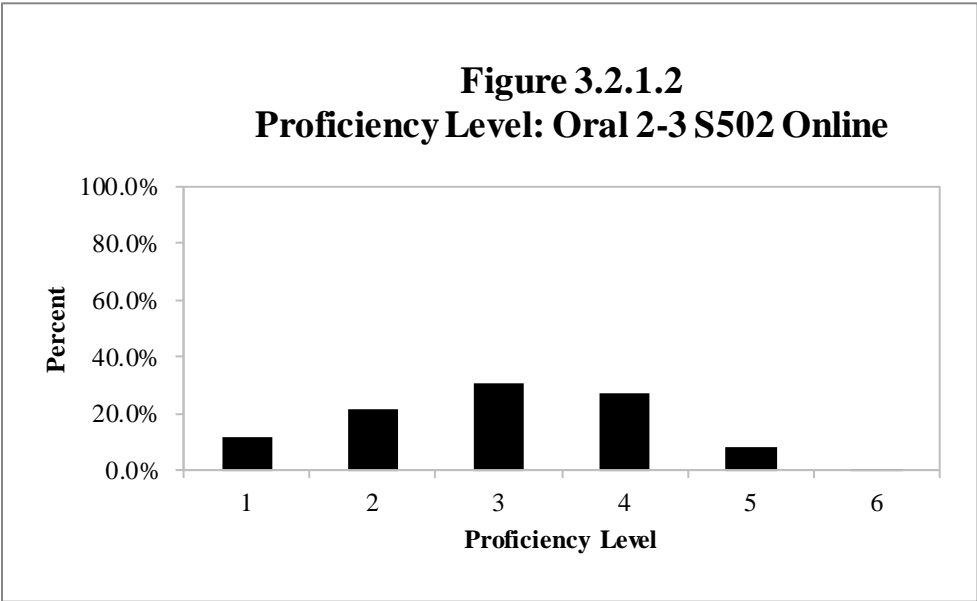


3.2.1.2 Grades 2–3

Table 3.2.1.2

Proficiency Level Distribution: Oral 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	14,105	12.34%	13,172	11.43%	27,277	11.89%
2	28,453	24.89%	21,190	18.39%	49,643	21.63%
3	35,708	31.24%	34,236	29.72%	69,944	30.48%
4	25,918	22.68%	36,328	31.53%	62,246	27.12%
5	9,254	8.10%	9,426	8.18%	18,680	8.14%
6	860	0.75%	851	0.74%	1,711	0.75%
Total	114,298	100.00%	115,203	100.00%	229,501	100.00%

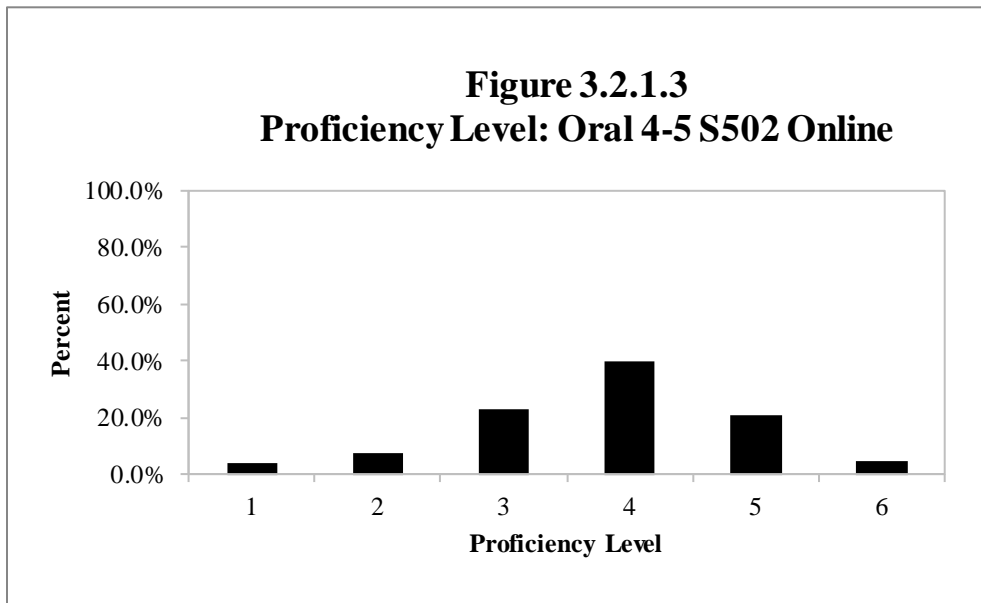


3.2.1.3 Grades 4–5

Table 3.2.1.3

Proficiency Level Distribution: Oral 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	3,830	3.24%	4,666	4.96%	8,496	4.01%
2	7,527	6.37%	8,005	8.52%	15,532	7.32%
3	26,801	22.70%	22,327	23.76%	49,128	23.17%
4	45,661	38.67%	39,080	41.58%	84,741	39.96%
5	27,378	23.19%	16,688	17.76%	44,066	20.78%
6	6,874	5.82%	3,216	3.42%	10,090	4.76%
Total	118,071	100.00%	93,982	100.00%	212,053	100.00%

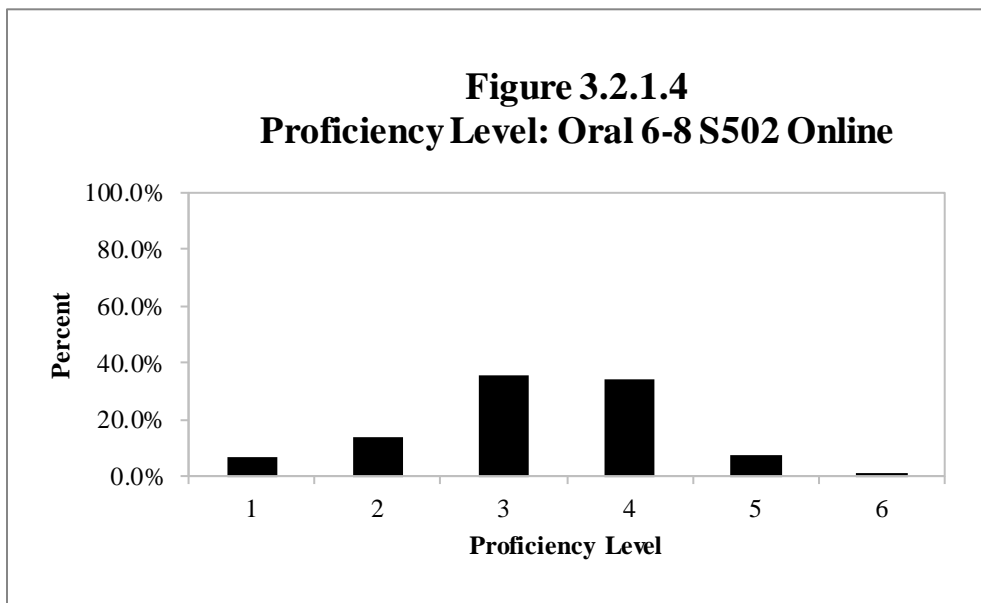


3.2.1.4 Grades 6–8

Table 3.2.1.4

Proficiency Level Distribution: Oral 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	3,911	5.14%	5,364	7.04%	6,434	9.34%	15,709	7.10%
2	9,528	12.53%	11,272	14.79%	10,211	14.83%	31,011	14.02%
3	28,503	37.48%	26,712	35.05%	24,478	35.55%	79,693	36.04%
4	27,837	36.60%	26,406	34.65%	21,599	31.37%	75,842	34.30%
5	5,666	7.45%	5,560	7.30%	5,259	7.64%	16,485	7.46%
6	603	0.79%	900	1.18%	876	1.27%	2,379	1.08%
Total	76,048	100.00%	76,214	100.00%	68,857	100.00%	221,119	100.00%

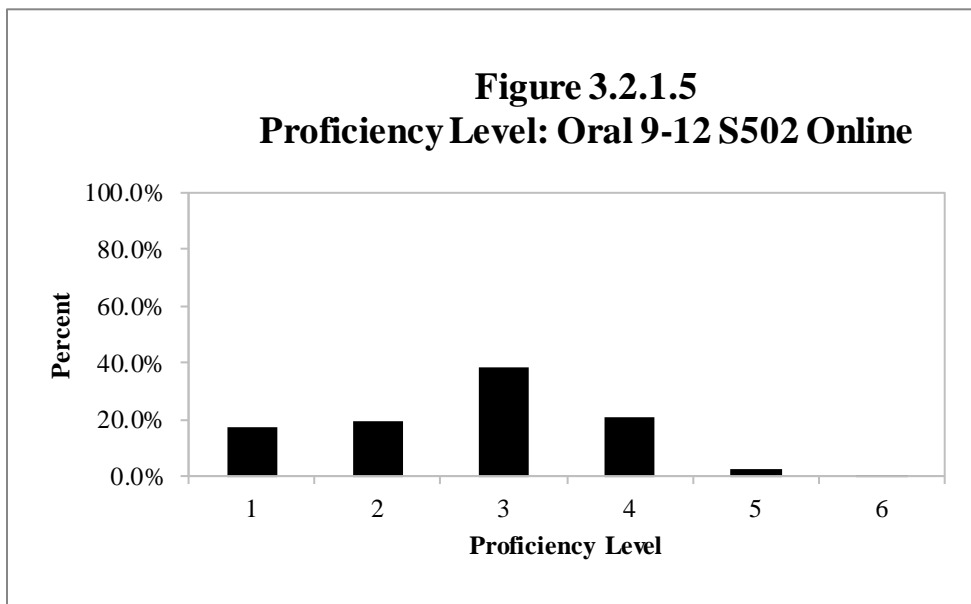


3.2.1.5 Grades 9-12

Table 3.2.1.5

Proficiency Level Distribution: Oral 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	8,763	15.14%	10,251	19.77%	7,490	17.83%	6,206	19.15%	32,710	17.76%
2	11,364	19.64%	10,064	19.41%	8,082	19.24%	6,154	18.99%	35,664	19.37%
3	21,988	38.00%	18,964	36.57%	16,652	39.64%	13,468	41.57%	71,072	38.60%
4	13,694	23.66%	10,829	20.88%	8,410	20.02%	5,755	17.76%	38,688	21.01%
5	1,784	3.08%	1,508	2.91%	1,169	2.78%	700	2.16%	5,161	2.80%
6	275	0.48%	238	0.46%	201	0.48%	118	0.36%	832	0.45%
Total	57,868	100.00%	51,854	100.00%	42,004	100.00%	32,401	100.00%	184,127	100.00%



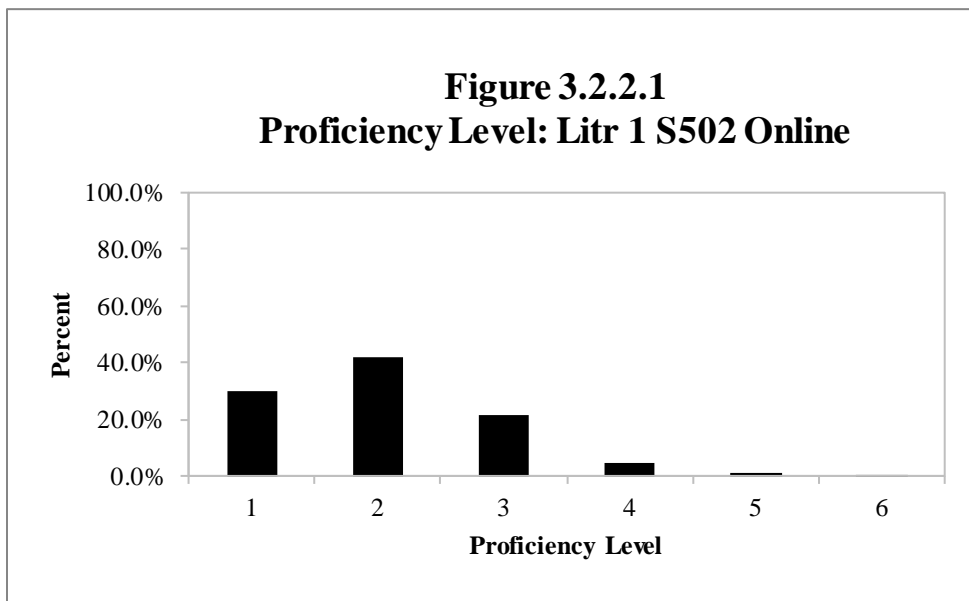
3.2.2 Literacy

3.2.2.1 Grade 1

Table 3.2.2.1

Proficiency Level Distribution: Litr 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	39,550	30.18%	39,550	30.18%
2	55,256	42.16%	55,256	42.16%
3	28,131	21.47%	28,131	21.47%
4	6,422	4.90%	6,422	4.90%
5	1,466	1.12%	1,466	1.12%
6	224	0.17%	224	0.17%
Total	131,049	100.00%	131,049	100.00%

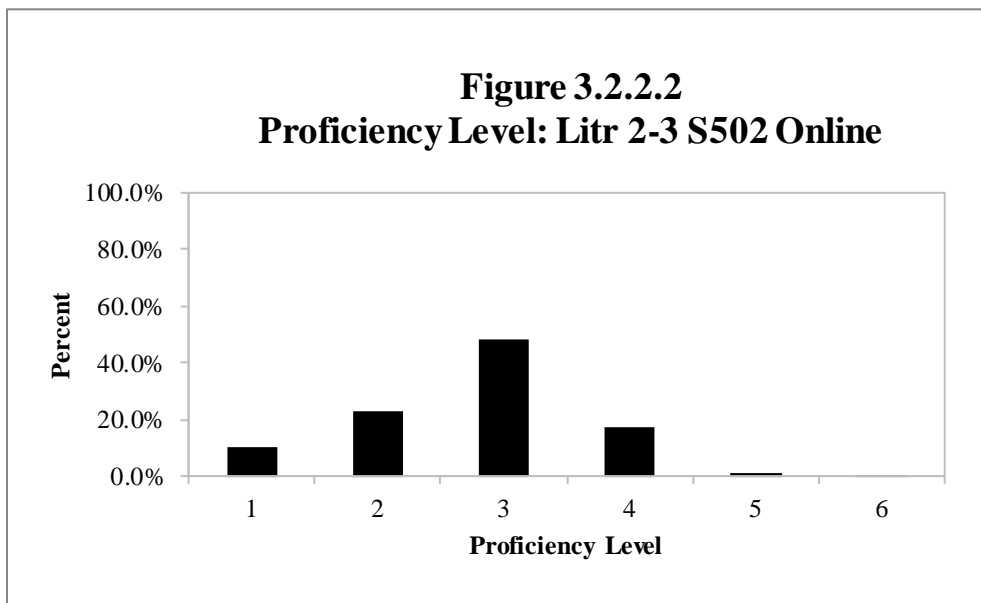


3.2.2.2 Grades 2–3

Table 3.2.2.2

Proficiency Level Distribution: Litr 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	13,435	10.85%	11,674	9.49%	25,109	10.17%
2	32,321	26.11%	23,836	19.37%	56,157	22.75%
3	59,107	47.75%	59,696	48.51%	118,803	48.13%
4	17,734	14.33%	25,311	20.57%	43,045	17.44%
5	1,117	0.90%	2,422	1.97%	3,539	1.43%
6	77	0.06%	124	0.10%	201	0.08%
Total	123,791	100.00%	123,063	100.00%	246,854	100.00%

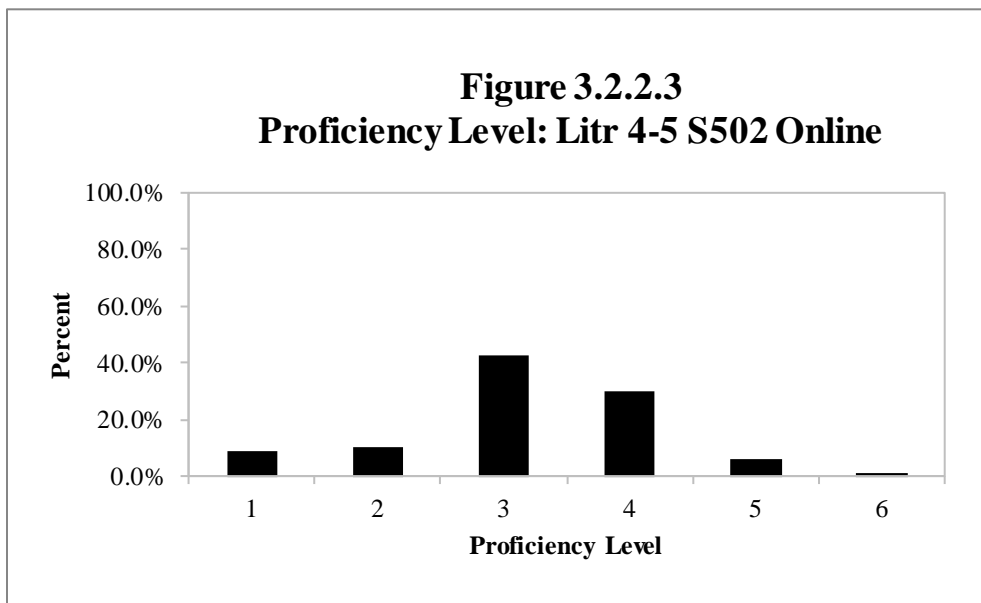


3.2.2.3 Grades 4–5

Table 3.2.2.3

Proficiency Level Distribution: Litr 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	10,902	8.91%	8,859	9.14%	19,761	9.01%
2	11,847	9.69%	10,502	10.83%	22,349	10.19%
3	52,768	43.14%	40,245	41.51%	93,013	42.42%
4	37,605	30.74%	29,091	30.01%	66,696	30.42%
5	7,536	6.16%	6,712	6.92%	14,248	6.50%
6	1,657	1.35%	1,534	1.58%	3,191	1.46%
Total	122,315	100.00%	96,943	100.00%	219,258	100.00%

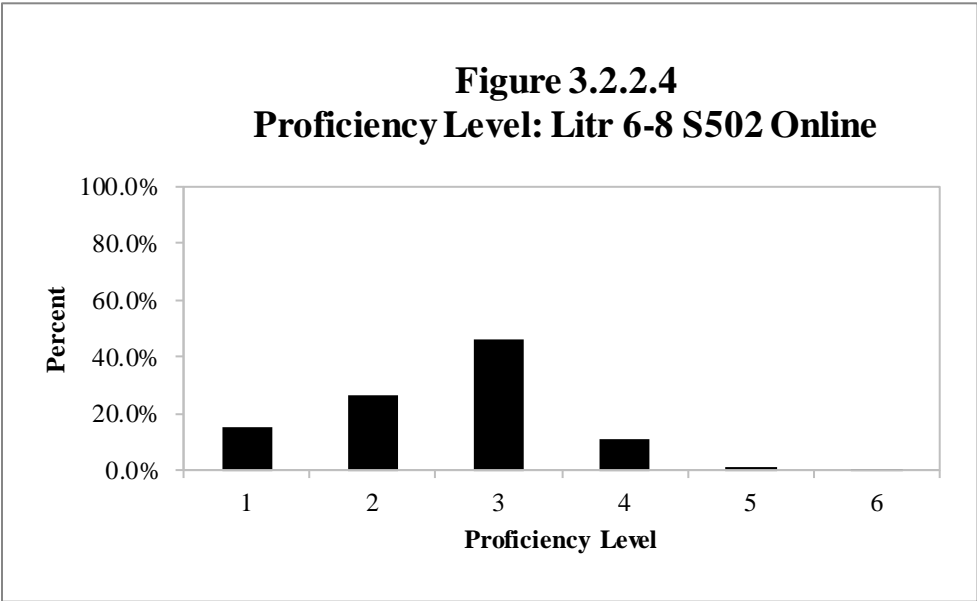


3.2.2.4 Grades 6–8

Table 3.2.2.4

Proficiency Level Distribution: Litr 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	11,206	14.10%	11,279	14.31%	11,867	16.80%	34,352	15.00%
2	22,340	28.10%	20,331	25.79%	17,826	25.24%	60,497	26.42%
3	39,313	49.46%	37,080	47.03%	29,406	41.63%	105,799	46.21%
4	6,181	7.78%	9,100	11.54%	10,368	14.68%	25,649	11.20%
5	398	0.50%	976	1.24%	1,114	1.58%	2,488	1.09%
6	53	0.07%	79	0.10%	49	0.07%	181	0.08%
Total	79,491	100.00%	78,845	100.00%	70,630	100.00%	228,966	100.00%

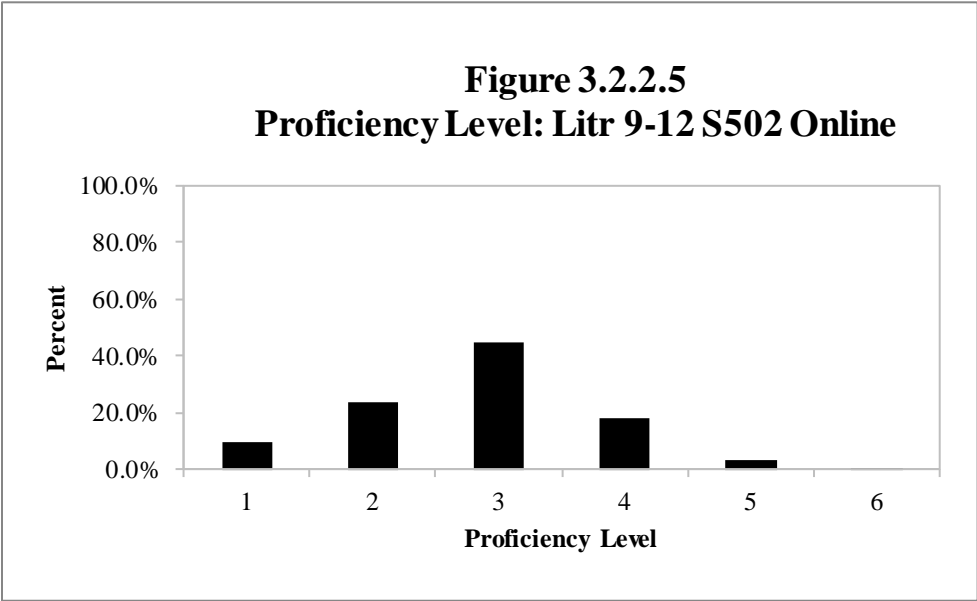


3.2.2.5 Grades 9-12

Table 3.2.2.5

Proficiency Level Distribution: Litr 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	5,294	8.91%	5,717	10.78%	4,183	9.70%	3,775	11.39%	18,969	10.05%
2	11,836	19.92%	12,472	23.51%	11,215	25.99%	9,600	28.97%	45,123	23.91%
3	27,890	46.94%	23,476	44.26%	18,741	43.44%	14,048	42.39%	84,155	44.59%
4	12,062	20.30%	9,423	17.76%	7,548	17.50%	4,954	14.95%	33,987	18.01%
5	2,193	3.69%	1,849	3.49%	1,415	3.28%	749	2.26%	6,206	3.29%
6	136	0.23%	109	0.21%	41	0.10%	16	0.05%	302	0.16%
Total	59,411	100.00%	53,046	100.00%	43,143	100.00%	33,142	100.00%	188,742	100.00%



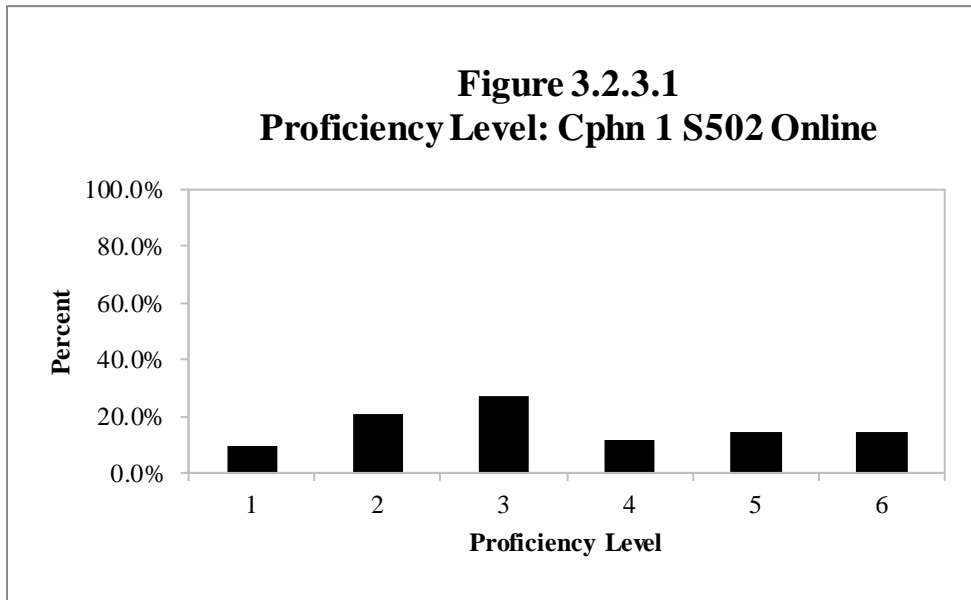
3.2.3 Comprehension

3.2.3.1 Grade 1

Table 3.2.3.1

Proficiency Level Distribution: Cphn 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	12,164	9.85%	12,164	9.85%
2	26,143	21.17%	26,143	21.17%
3	33,982	27.52%	33,982	27.52%
4	14,534	11.77%	14,534	11.77%
5	18,387	14.89%	18,387	14.89%
6	18,279	14.80%	18,279	14.80%
Total	123,489	100.00%	123,489	100.00%

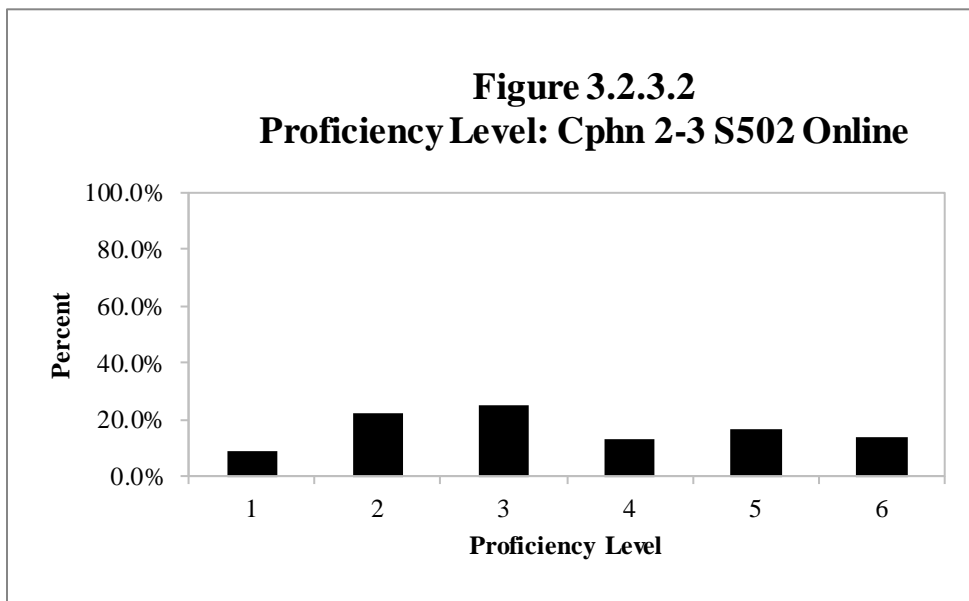


3.2.3.2 Grades 2–3

Table 3.2.3.2

Proficiency Level Distribution: Cphn 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	7,725	6.62%	13,090	11.23%	20,815	8.93%
2	26,571	22.78%	25,656	22.02%	52,227	22.40%
3	31,020	26.59%	27,321	23.45%	58,341	25.02%
4	17,046	14.61%	13,535	11.62%	30,581	13.12%
5	19,381	16.62%	19,057	16.35%	38,438	16.49%
6	14,896	12.77%	17,869	15.33%	32,765	14.05%
Total	116,639	100.00%	116,528	100.00%	233,167	100.00%

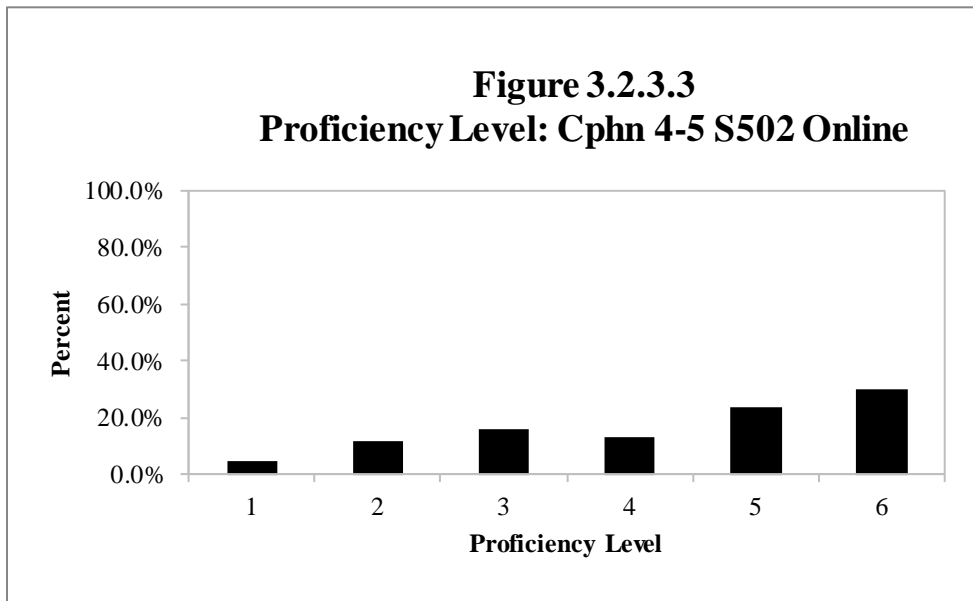


3.2.3.3 Grades 4–5

Table 3.2.3.3

Proficiency Level Distribution: Cphn 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	4,098	3.45%	6,005	6.37%	10,103	4.74%
2	13,675	11.53%	11,784	12.50%	25,459	11.96%
3	18,695	15.76%	15,629	16.57%	34,324	16.12%
4	15,301	12.90%	13,552	14.37%	28,853	13.55%
5	29,166	24.59%	21,468	22.77%	50,634	23.78%
6	37,689	31.77%	25,860	27.42%	63,549	29.85%
Total	118,624	100.00%	94,298	100.00%	212,922	100.00%

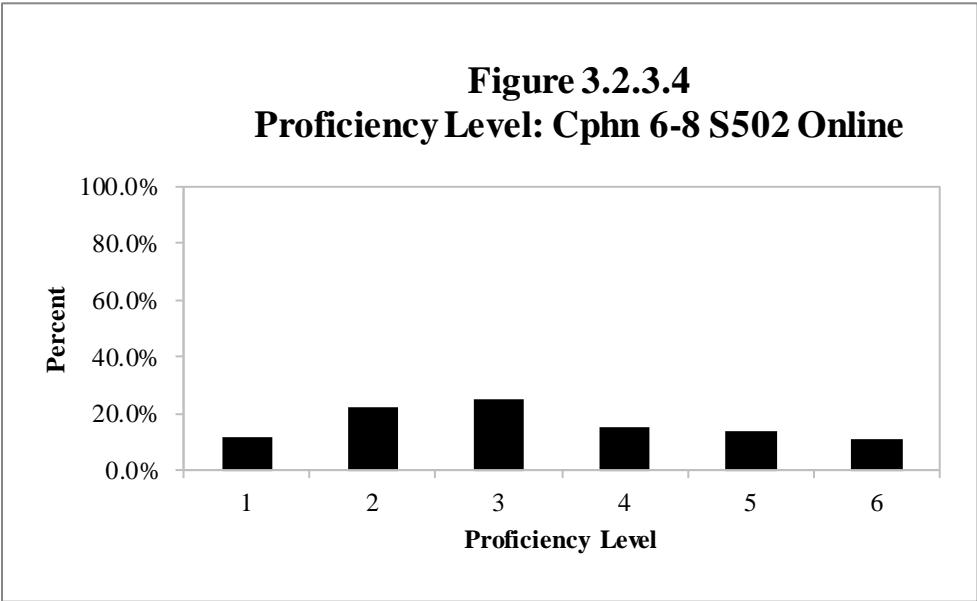


3.2.3.4 Grades 6–8

Table 3.2.3.4

Proficiency Level Distribution: Cphn 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	7,257	9.43%	9,234	12.01%	9,942	14.34%	26,433	11.84%
2	18,576	24.13%	16,678	21.70%	15,413	22.24%	50,667	22.70%
3	21,109	27.42%	19,273	25.07%	15,505	22.37%	55,887	25.04%
4	13,143	17.07%	12,152	15.81%	9,303	13.42%	34,598	15.50%
5	10,962	14.24%	10,103	13.14%	9,283	13.39%	30,348	13.60%
6	5,950	7.73%	9,427	12.26%	9,864	14.23%	25,241	11.31%
Total	76,997	100.00%	76,867	100.00%	69,310	100.00%	223,174	100.00%

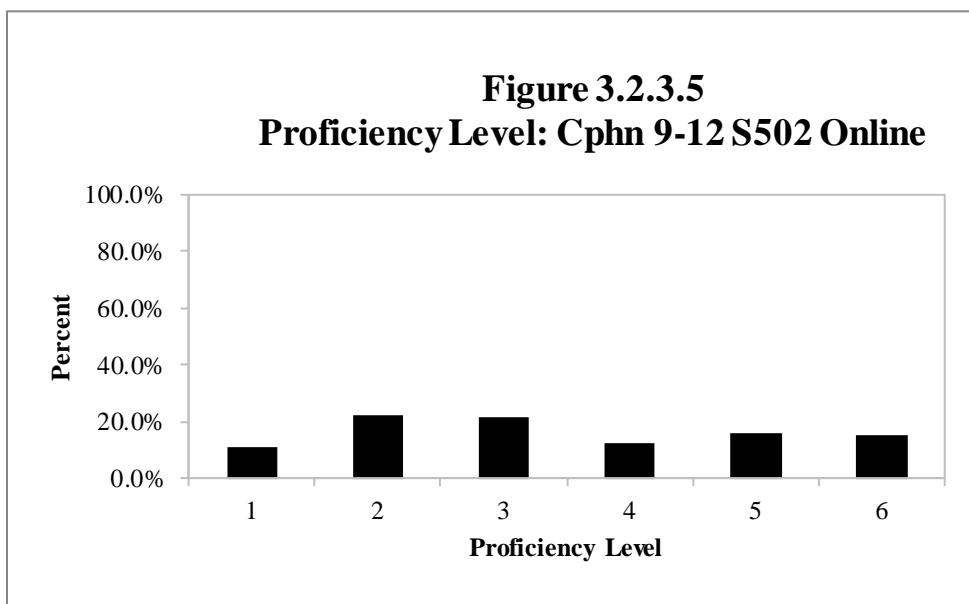


3.2.3.5 Grades 9-12

Table 3.2.3.5

Proficiency Level Distribution: Cphn 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	5,345	9.26%	6,349	12.25%	5,065	11.99%	4,210	12.98%	20,969	11.38%
2	12,674	21.96%	11,926	23.01%	9,605	22.75%	7,624	23.51%	41,829	22.71%
3	12,996	22.52%	11,116	21.45%	8,756	20.74%	6,772	20.88%	39,640	21.52%
4	7,893	13.68%	6,472	12.49%	5,248	12.43%	4,107	12.67%	23,720	12.88%
5	9,731	16.86%	8,059	15.55%	6,893	16.32%	5,474	16.88%	30,157	16.37%
6	9,071	15.72%	7,908	15.26%	6,661	15.77%	4,240	13.08%	27,880	15.14%
Total	57,710	100.00%	51,830	100.00%	42,228	100.00%	32,427	100.00%	184,195	100.00%



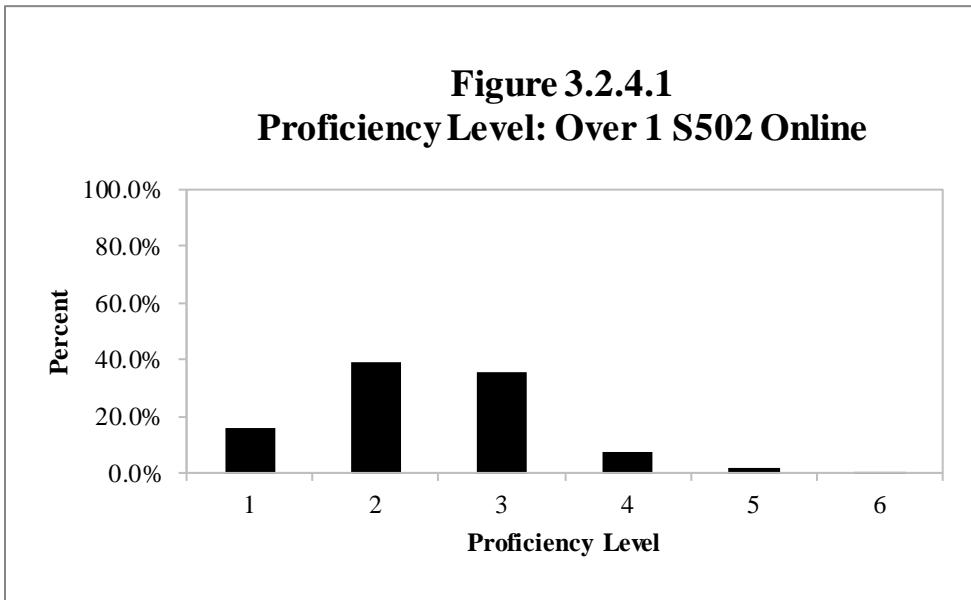
3.2.4 Overall

3.2.4.1 Grade 1

Table 3.2.4.1

Proficiency Level Distribution: Over 1 S502 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	18,792	16.12%	18,792	16.12%
2	45,534	39.06%	45,534	39.06%
3	41,277	35.40%	41,277	35.40%
4	8,829	7.57%	8,829	7.57%
5	1,939	1.66%	1,939	1.66%
6	215	0.18%	215	0.18%
Total	116,586	100.00%	116,586	100.00%

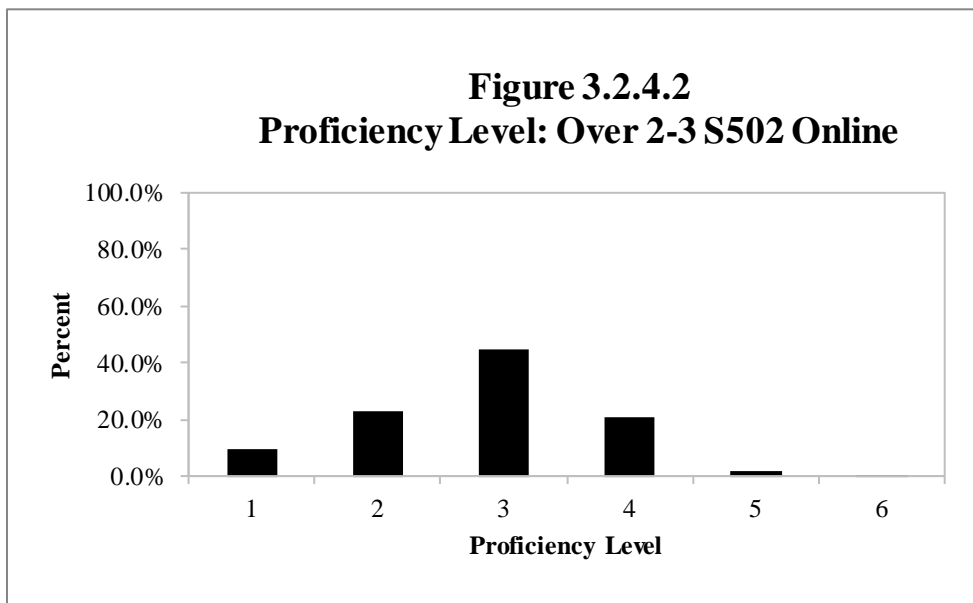


3.2.4.2 Grades 2–3

Table 3.2.4.2

Proficiency Level Distribution: Over 2-3 S502 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	10,997	10.07%	10,203	9.27%	21,200	9.67%
2	29,238	26.76%	20,733	18.83%	49,971	22.78%
3	48,762	44.64%	49,865	45.29%	98,627	44.97%
4	18,464	16.90%	26,797	24.34%	45,261	20.64%
5	1,713	1.57%	2,454	2.23%	4,167	1.90%
6	67	0.06%	41	0.04%	108	0.05%
Total	109,241	100.00%	110,093	100.00%	219,334	100.00%

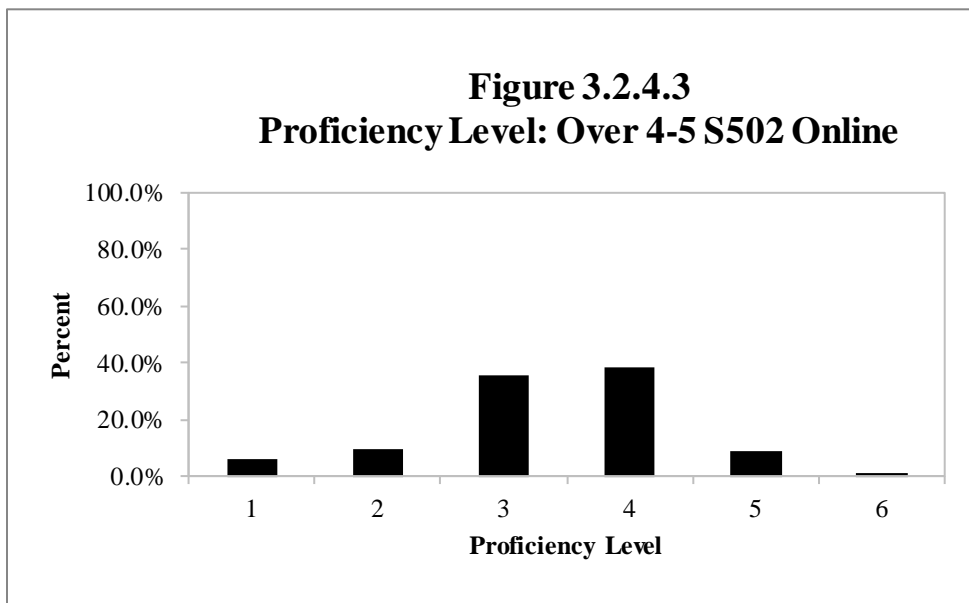


3.2.4.3 Grades 4–5

Table 3.2.4.3

Proficiency Level Distribution: Over 4-5 S502 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	6,142	5.63%	5,759	6.60%	11,901	6.06%
2	9,929	9.10%	9,017	10.34%	18,946	9.65%
3	38,563	35.32%	31,742	36.40%	70,305	35.80%
4	42,625	39.05%	32,533	37.30%	75,158	38.27%
5	10,394	9.52%	7,221	8.28%	17,615	8.97%
6	1,515	1.39%	937	1.07%	2,452	1.25%
Total	109,168	100.00%	87,209	100.00%	196,377	100.00%

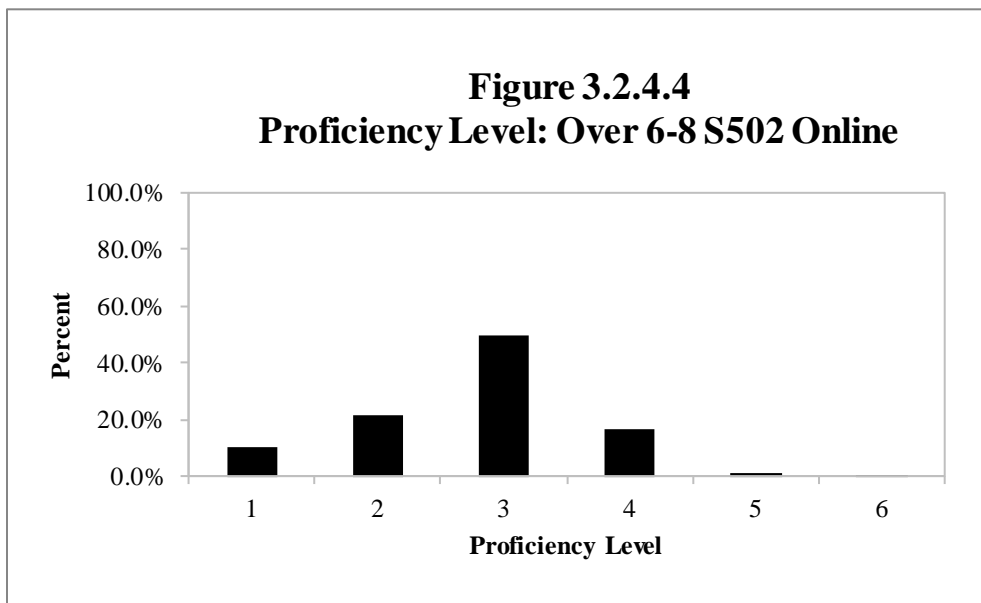


3.2.4.4 Grades 6–8

Table 3.2.4.4

Proficiency Level Distribution: Over 6-8 S502 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	5,839	8.28%	7,081	10.07%	7,874	12.38%	20,794	10.17%
2	15,615	22.15%	14,676	20.88%	13,438	21.12%	43,729	21.39%
3	38,657	54.84%	34,689	49.35%	28,633	45.00%	101,979	49.89%
4	9,772	13.86%	12,659	18.01%	12,299	19.33%	34,730	16.99%
5	552	0.78%	1,098	1.56%	1,319	2.07%	2,969	1.45%
6	55	0.08%	90	0.13%	61	0.10%	206	0.10%
Total	70,490	100.00%	70,293	100.00%	63,624	100.00%	204,407	100.00%

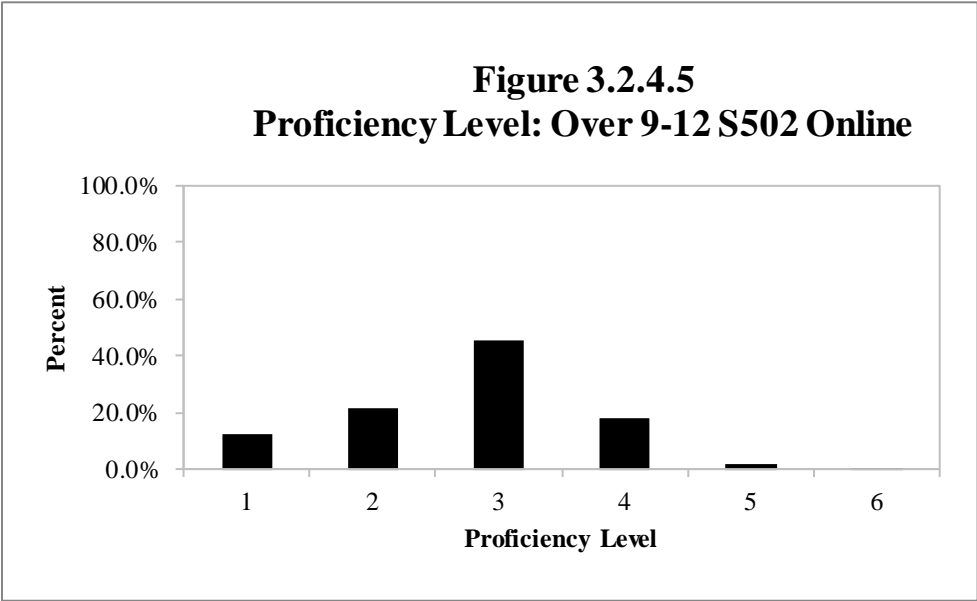


3.2.4.5 Grades 9-12

Table 3.2.4.5

Proficiency Level Distribution: Over 9-12 S502 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	5,517	10.35%	6,671	13.90%	4,789	12.30%	3,982	13.23%	20,959	12.30%
2	10,167	19.07%	10,173	21.19%	8,846	22.71%	7,869	26.15%	37,055	21.75%
3	25,291	47.45%	21,072	43.89%	17,498	44.93%	13,232	43.97%	77,093	45.26%
4	10,880	20.41%	8,859	18.45%	6,911	17.74%	4,525	15.04%	31,175	18.30%
5	1,366	2.56%	1,178	2.45%	880	2.26%	475	1.58%	3,899	2.29%
6	84	0.16%	57	0.12%	23	0.06%	7	0.02%	171	0.10%
Total	53,305	100.00%	48,010	100.00%	38,947	100.00%	30,090	100.00%	170,352	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 the document *A State's Guidance to the U.S. Department of Education's Assessment Peer Review Process* (U.S. 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, and is linked to the Assessment User Argument (AUA) to support the claims of validity of the Online ACCESS assessment. WIDA structures its validity arguments using the 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 the WIDA AUA 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 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 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

Important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test. Administration and scoring may also be relevant to content-based evidence. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on test content can also come from expert judgment of the relationship between parts of the test and content. Section 4.2.2, Dimensionality and Content Knowledge in ACCESS Tests, addresses the validity of dimensionality of ACCESS Online tests in relation to content knowledge by ability level.

4.1.2 Response Processes

Theoretical and empirical analyses of the response processes of test-takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test-takers. Evidence based on response processes generally comes from analysis of individual responses. Evidence of response processes can contribute to answering questions about differences in meaning or interpretation of test scores across relevant subgroups of test-takers. Studies of response processes are not limited to the test-taker. Assessment often relies on observers or judges to record and/or evaluate test-takers' performance or products.

4.1.3 Internal Structure

Analyses of the internal structure of a test can indicate the degree to which the relationships among the test items and test components conform to the construct on which the proposed test score interpretations are based. The conceptual framework for a test may imply a single dimension of behavior, or it may posit several components that are each expected to be homogeneous.

4.1.4 Relation to Other Variables

In many cases, the intended interpretation for a given use implies that the construct should be related to some other variables, and as a result, analysis of the relationship of the scores to variables external to the test provides another important source of validity evidence. Evidence about relation to other variables is also used to investigate questions of differential prediction for subgroups. In the test-criterion relationship, the fundamental question is the accuracy with which test scores predict criterion performance. Historically, two designs, often called predictive and concurrent, have been differentiated for evaluating test-criterion relationships. A predictive study

indicates the strength of the relationship between test scores and criterion scores that are obtained at a later time. A concurrent study obtains test scores and criterion information at about the same time. Section 4.2.1, *Enhancement of ACCESS Online Tests in Comparison with ACCESS Paper Tests*, addresses how multistage adaptivity in Online ACCESS tests improved precision compared to Paper ACCESS tests.

4.2 Annual Validity Studies

4.2.1 Enhancement of ACCESS Online Tests in Comparison with ACCESS Paper Tests

How can one construct a test that provides accurate measurements across the range of performance levels while providing adequate coverage of all of the critical areas of the domain and not be unmanageably long? MacGregor, Yen, and Yu (2021) discussed the approach taken in a linear test of academic English language and how the transition to a computer-based test allowed for a design that better fit the demands of the test. It also described the multistage adaptive approach that was devised. This approach allows for a test that covers a broad range of performance levels while including items that assess the language of the content areas as described in the ELD Standards underpinning the test. The design also allows for a test that is closely tailored to the ability level of the ELs taking the test and that therefore produces a more precise measure. The efficacy of the design in enhancing measurement of two versions of a high-stakes English language assessment is explored, and the implications of the results are discussed.

4.2.2 Dimensionality and Content Knowledge in ACCESS Tests

The study of Min, Bishop, and Cook (2021) explored the interplay between content knowledge and reading ability in a large-scale multistage adaptive English for academic purposes (EAP) reading assessment at a range of ability levels across 1st to 12th graders. The datasets for this study were item-level responses to the reading tests of ACCESS for ELLs Online 2.0. A sample of 10,000 test-takers were randomly drawn from the test-taking population at five grade clusters, first without manipulation of proficiency levels and then with manipulation of proficiency levels. The results indicated that although the bifactor multidimensional item response theory model fit the data significantly better than the unidimensional two-parameter logistic model for Grade 1, no clear evidence could be found regarding the dimensionality of the test for Grades 2 to 12. However, content knowledge was consistently found to contribute substantially to test performance for low-ability-level test-takers across all grade clusters. The findings indicate that EAP reading ability is a multidimensional construct in the onset of EAP reading ability development, but the presence of multidimensionality decreases as proficiency level and grade level increase. This study provides insights into the developmental pattern of the interplay between language and content in EAP reading contexts.

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. We report the **interrater agreement** rates that the raters achieved when evaluating students' responses to the Writing and Speaking tasks in Section 5.2. 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 have provided the results in Section 5.5. In addition, we examine the CSEM of these scores in Section 5.6. 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 good 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.

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). We report the **interrater agreement rates** for scoring students' responses to

the Writing and Speaking tasks in Section 5.2. These values reflect how consistent the students' scores would be if different groups of raters scored their responses, while we present a detailed description of the methods, data sources, and procedures in this section.

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 measure 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 **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 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

$$\bar{\rho} = \frac{\sigma_{\theta}^2 - \text{average}(CSEM_{observed}^2)}{\sigma_{\theta}^2}$$

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 the 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). 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 scale scores for all students ranged from 0.82 to 0.87 across the grade-level clusters. The Rasch student separation reliability coefficients ranged from 0.82 to 0.87 for male students; 0.81 to 0.86 for female students; 0.82 to 0.87 for Hispanic students; 0.80 to 0.86 for non-Hispanic students; and 0.81 to 0.89 for students with an IEP.

For Reading, the Rasch student separation reliability coefficients based on the domain scale scores for all students ranged from 0.87 to 0.91 across the grade-level clusters. The Rasch student separation reliability coefficients ranged from 0.87 to 0.91 for male students; 0.87 to 0.90 for female students; 0.85 to 0.90 for Hispanic students; 0.88 to 0.91 for non-Hispanic students; and 0.84 to 0.88 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

$$SEM = SD\sqrt{1 - reliability}$$

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 95% probability of falling within the band extending from the observed score minus 2 SEM to the observed score plus 2 SEM. 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 (10 minus 2 or plus 2 SEMs). Similarly, if one SEM equals 1 score point, we would say with 95% certainty that the student's true score lies between 9 and 11. The smaller the value of 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 in Speaking are 0 to 6 for pre-A, 0 to 18 for Tier A, and 0 to 24 for Tier B/C. As described in Section 2.3, because of the semiadaptive nature of the Speaking test, Speaking Tier B/C students are awarded 2 full points on all three PL 1 tasks that they did not take, such that the officially reported total raw score points for the Speaking Tier B/C form range from 6 to 30. However, since the Cronbach's alpha for the Speaking Tier B/C form was computed using students' raw scores on the six Speaking tasks the students actually took, the total possible raw score points reported in 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, and 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 0.80 to 0.90. The Writing Tier A Cronbach's coefficient alphas ranged from 0.81 to 0.90 for male students; 0.79 to 0.89 for female students; 0.80 to 0.90 for Hispanic students; 0.79 to 0.89 for non-Hispanic students; and 0.81 to 0.89 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 0.69 to 0.82. The Writing Tier B/C Cronbach's coefficient alphas ranged from 0.69 to 0.83 for male students; 0.67 to 0.79 for female students; 0.68 to 0.82 for Hispanic students; 0.68 to 0.79 for non-Hispanic students; and 0.67 to 0.86 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 0.84 to 0.85. The Cronbach's coefficient alphas ranged from 0.83 to 0.86 for male students; 0.83 to 0.86 for female students; 0.83 to 0.85 for Hispanic students; 0.81 to 0.87 for non-Hispanic students; and 0.81 to 0.89 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 0.78 to 0.82. The Cronbach's coefficient alphas ranged from 0.78 to 0.82 for male students; 0.78 to 0.82 for female students; 0.79 to 0.83 for Hispanic students; 0.76 to 0.80 for non-Hispanic students; and 0.76 to 0.83 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 0.82 to 0.85. The Cronbach's coefficient alphas ranged from 0.81 to 0.85 for male students; 0.82 to 0.85 for female students; 0.82 to 0.85 for Hispanic students; 0.81 to 0.83 for non-Hispanic students; and 0.81 to 0.85 for students with an IEP.

5.1.1 Listening

Table 5.1.1.1

Reliabilities of Domain Scores: List S502 Online

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	128,188	54	0.87
2-3	244,652	54	0.86
4-5	225,824	54	0.82
6-8	236,488	54	0.84
9-12	195,881	54	0.87

Table 5.1.1.2

Reliabilities of Domain Scores: List S502 Online by Gender

Cluster	No. of Items	Female		Male	
		No. of Students	Rasch Student Separation Reliability Coefficient	No. of Students	Rasch Student Separation Reliability Coefficient
1	54	59,975	0.86	66,774	0.87
2-3	54	113,452	0.85	128,705	0.86
4-5	54	101,816	0.81	121,608	0.82
6-8	54	102,082	0.84	131,335	0.85
9-12	54	85,407	0.86	107,754	0.87

Table 5.1.1.3

Reliabilities of Domain Scores: List S502 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	54	82,907	0.87	40,670	0.86
2-3	54	161,936	0.86	74,279	0.84
4-5	54	154,752	0.82	59,737	0.80
6-8	54	165,828	0.84	57,821	0.83
9-12	54	132,180	0.87	53,279	0.85

Table 5.1.1.4

Reliabilities of Domain Scores: List S502 Online by IEP Status

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	10,185	54	0.89
2-3	22,046	54	0.88
4-5	27,186	54	0.82
6-8	36,152	54	0.81
9-12	25,836	54	0.82

5.1.2 Reading

Table 5.1.2.1

Reliabilities of Domain Scores: Read S502 Online

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	131,078	72	0.88
2-3	246,934	72	0.87
4-5	225,166	72	0.90
6-8	236,259	72	0.90
9-12	193,768	72	0.91

Table 5.1.2.2

Reliabilities of Domain Scores: Read S502 Online by Gender

Cluster	No. of Items	Female		Male	
		No. of Students	Rasch Student Separation Reliability Coefficient	No. of Students	Rasch Student Separation Reliability Coefficient
1	72	61,063	0.89	68,520	0.88
2-3	72	114,188	0.87	130,223	0.87
4-5	72	100,957	0.89	121,787	0.90
6-8	72	101,408	0.90	131,755	0.91
9-12	72	84,193	0.90	106,944	0.91

Table 5.1.2.3

Reliabilities of Domain Scores: Read S502 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	84,974	0.85	41,425	0.90
2-3	72	163,626	0.86	74,856	0.88
4-5	72	154,398	0.89	59,469	0.90
6-8	72	165,963	0.90	57,532	0.91
9-12	72	131,208	0.90	52,394	0.91

Table 5.1.2.4

Reliabilities of Domain Scores: Read S502 Online by IEP Status

Cluster	No. of Students	No. of Items	Rasch Student Separation Reliability Coefficient
1	10,505	72	0.84
2-3	22,425	72	0.84
4-5	27,426	72	0.88
6-8	36,299	72	0.87
9-12	25,590	72	0.87

5.1.3 Writing

Table 5.1.3.1

Reliabilities of Domain Scores: Writ S502 Online

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	A	115,655	2	0-18	0.80	1.15
	B/C	21,518	2	0-18	0.69	1.28
2-3	A	73,892	2	0-18	0.89	1.06
	B/C	187,337	2	0-18	0.82	0.96
4-5	A	46,022	2	0-18	0.90	1.02
	B/C	188,043	2	0-18	0.70	1.16
6-8	A	88,587	2	0-18	0.89	0.95
	B/C	155,986	2	0-18	0.71	1.03
9-12	A	63,380	2	0-18	0.87	1.12
	B/C	138,719	2	0-18	0.69	1.17

Table 5.1.3.2

Reliabilities of Domain Scores: Writ S502 Online by Gender

Cluster	Tier	No. of Tasks	Total Possible Raw Score Points	Female			Male		
				No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	A	2	0-18	53,011	0.79	1.16	61,359	0.81	1.14
	B/C	2	0-18	10,909	0.68	1.28	10,341	0.70	1.29
2-3	A	2	0-18	31,273	0.89	1.06	41,677	0.90	1.06
	B/C	2	0-18	89,751	0.79	0.93	95,901	0.83	0.98
4-5	A	2	0-18	18,350	0.89	1.02	27,002	0.90	1.01
	B/C	2	0-18	86,770	0.67	1.15	99,434	0.72	1.17
6-8	A	2	0-18	35,140	0.88	0.96	52,166	0.89	0.94
	B/C	2	0-18	69,693	0.68	1.02	84,275	0.72	1.03
9-12	A	2	0-18	25,859	0.86	1.12	36,372	0.87	1.13
	B/C	2	0-18	61,775	0.67	1.16	75,311	0.69	1.18

Table 5.1.3.3

Reliabilities of Domain Scores: Writ S502 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	79,709	0.80	1.15	31,910	0.79	1.15
	B/C	2	0-18	8,921	0.70	1.29	11,749	0.68	1.27
2-3	A	2	0-18	54,492	0.90	1.07	16,740	0.89	1.05
	B/C	2	0-18	118,320	0.82	0.97	62,781	0.79	0.93
4-5	A	2	0-18	33,483	0.90	1.01	9,565	0.88	1.05
	B/C	2	0-18	127,050	0.70	1.15	52,344	0.69	1.18
6-8	A	2	0-18	65,040	0.89	0.94	18,192	0.85	0.98
	B/C	2	0-18	106,587	0.71	1.02	41,525	0.72	1.05
9-12	A	2	0-18	45,276	0.87	1.12	13,715	0.83	1.12
	B/C	2	0-18	91,287	0.68	1.16	41,202	0.70	1.19

Table 5.1.3.4

Reliabilities of Domain Scores: Writ S502 Online by IEP Status

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	A	10,167	2	0-18	0.82	1.11
	B/C	806	2	0-18	0.76	1.28
2-3	A	11,607	2	0-18	0.89	1.09
	B/C	12,034	2	0-18	0.86	1.05
4-5	A	10,850	2	0-18	0.87	1.04
	B/C	17,505	2	0-18	0.75	1.18
6-8	A	19,488	2	0-18	0.84	0.93
	B/C	18,002	2	0-18	0.73	1.04
9-12	A	9,708	2	0-18	0.81	1.12
	B/C	17,016	2	0-18	0.67	1.14

5.1.4 Speaking

Table 5.1.4.1

Reliabilities of Domain Scores: Spek S502 Online

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	Pre-A	5,634	3	0-6	0.85	0.77
	A	55,599	6	0-18	0.81	1.32
	B/C	67,181	6	0-24	0.82	1.62
2-3	Pre-A	9,151	3	0-6	0.85	0.64
	A	68,806	6	0-18	0.78	1.36
	B/C	165,709	6	0-24	0.83	1.56
4-5	Pre-A	3,237	3	0-6	0.84	0.80
	A	29,624	6	0-18	0.81	1.37
	B/C	192,651	6	0-24	0.83	1.52
6-8	Pre-A	5,764	3	0-6	0.85	0.72
	A	46,706	6	0-18	0.82	1.29
	B/C	183,296	6	0-24	0.84	1.47
9-12	Pre-A	11,988	3	0-6	0.84	0.65
	A	72,435	6	0-18	0.81	1.27
	B/C	110,686	6	0-24	0.85	1.40

Table 5.1.4.2

Reliabilities of Domain Scores: Spek S502 Online by Gender

Cluster	Tier	No. of Tasks	Total Possible Raw Score Points	Female			Male		
				No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	Pre-A	3	0-6	2,179	0.85	0.73	3,380	0.85	0.79
	A	6	0-18	24,374	0.80	1.32	30,572	0.81	1.32
	B/C	6	0-24	33,554	0.82	1.62	32,887	0.81	1.61
2-3	Pre-A	3	0-6	3,709	0.86	0.61	5,283	0.84	0.65
	A	6	0-18	29,790	0.78	1.36	38,199	0.78	1.36
	B/C	6	0-24	79,788	0.83	1.56	84,409	0.83	1.57
4-5	Pre-A	3	0-6	1,331	0.85	0.79	1,814	0.83	0.81
	A	6	0-18	12,102	0.81	1.37	17,104	0.81	1.36
	B/C	6	0-24	88,202	0.83	1.52	102,568	0.83	1.51
6-8	Pre-A	3	0-6	2,316	0.84	0.71	3,316	0.86	0.73
	A	6	0-18	18,626	0.82	1.30	27,422	0.82	1.28
	B/C	6	0-24	80,338	0.85	1.47	100,624	0.84	1.46
9-12	Pre-A	3	0-6	4,895	0.83	0.64	6,813	0.85	0.64
	A	6	0-18	30,159	0.80	1.28	41,170	0.82	1.26
	B/C	6	0-24	49,684	0.85	1.41	59,698	0.85	1.39

Table 5.1.4.3

Reliabilities of Domain Scores: Spek S502 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	4,126	0.85	0.78	1,255	0.86	0.74
	A	6	0-18	39,164	0.81	1.32	14,525	0.80	1.32
	B/C	6	0-24	39,655	0.82	1.60	25,118	0.81	1.63
2-3	Pre-A	3	0-6	6,729	0.85	0.64	2,011	0.81	0.61
	A	6	0-18	50,137	0.79	1.36	16,261	0.76	1.38
	B/C	6	0-24	104,352	0.84	1.55	55,812	0.82	1.58
4-5	Pre-A	3	0-6	2,236	0.83	0.82	554	0.82	0.72
	A	6	0-18	21,497	0.81	1.37	6,207	0.76	1.36
	B/C	6	0-24	130,981	0.83	1.50	52,775	0.82	1.54
6-8	Pre-A	3	0-6	4,286	0.85	0.73	843	0.83	0.67
	A	6	0-18	34,167	0.83	1.29	9,619	0.78	1.27
	B/C	6	0-24	127,167	0.85	1.46	46,918	0.83	1.49
9-12	Pre-A	3	0-6	8,882	0.83	0.65	2,029	0.87	0.53
	A	6	0-18	50,831	0.82	1.27	17,427	0.77	1.26
	B/C	6	0-24	72,098	0.85	1.39	33,596	0.83	1.41

Table 5.1.4.4

Reliabilities of Domain Scores: Spek S502 Online by IEP Status

Cluster	Tier	No. of Students	No. of Tasks	Total Possible Raw Score Points	Cronbach's Alpha	SEM
1	Pre-A	934	3	0-6	0.84	0.80
	A	5,931	6	0-18	0.83	1.34
	B/C	3,442	6	0-24	0.81	1.64
2-3	Pre-A	1,789	3	0-6	0.85	0.57
	A	10,291	6	0-18	0.78	1.34
	B/C	10,043	6	0-24	0.84	1.57
4-5	Pre-A	346	3	0-6	0.81	0.73
	A	7,151	6	0-18	0.76	1.33
	B/C	19,732	6	0-24	0.83	1.53
6-8	Pre-A	556	3	0-6	0.87	0.60
	A	10,385	6	0-18	0.80	1.22
	B/C	25,229	6	0-24	0.84	1.45
9-12	Pre-A	1,065	3	0-6	0.89	0.56
	A	12,518	6	0-18	0.82	1.22
	B/C	12,318	6	0-24	0.85	1.38

5.2 Interrater Agreement Rates

The tables below provide information on interrater agreement of DRC raters who scored a sample of 20% of the students' responses on the Online Speaking and Writing tests. We describe the details about the scoring of performance tasks in Part II, Section 3.2.2. These tables show, for each of the tasks, the percentage of agreement between two raters who independently scored students' responses to that task. The first column shows the task, and the second column shows the number of responses that raters double scored. DRC selects a sample of 20% of all responses scored, chosen at random during the operational scoring process, for double scoring. The next columns show the rates of agreement in the scores that the raters assigned.

For Writing, the scoring rubric that the raters used 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 scoring rubric that the raters used 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 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%. DRC defines this “**agreement**” as being scored as adjacent agreement (AD) for Writing and exact agreement (EX) for speaking. 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 97%. For Speaking, the lowest interrater agreement rate was 77%.

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 S502 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	72,242	99	1	0
	2	68,566	99	1	0

Table 5.2.3.1.2

Interrater Agreement: Writ 1 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	9,908	100	0	0
	2	10,460	99	1	0

5.2.3.2 Grade 2–3

Table 5.2.3.2.1

Interrater Agreement: Writ 2-3 A S502 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	52,602	98	2	0
	2	48,988	98	2	0

Table 5.2.3.2.2

Interrater Agreement: Writ 2-3 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	93,934	98	2	0
	2	95,048	99	1	0

5.2.3.3 Grades 4–5

Table 5.2.3.3.1

Interrater Agreement: Writ 4-5 A S502 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	3,228	99	1	0
		KB	18,302	98	2	0
	2	HW	3,354	99	1	0
		KB	18,334	98	2	0

Table 5.2.3.3.2

Interrater Agreement: Writ 4-5 B/C S502 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	8,982	99	1	0
		KB	76,414	98	2	0
	2	HW	8,148	98	2	0
		KB	80,084	98	2	0

5.2.3.4 Grades 6–8

Table 5.2.3.4.1

Interrater Agreement: Writ 6-8 A S502 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	194	98	2	0
		KB	37,356	99	1	0
	2	HW	184	98	2	0
		KB	37,350	98	2	0

Table 5.2.3.4.2

Interrater Agreement: Writ 6-8 B/C S502 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	226	100	0	0
		KB	67,622	99	1	0
	2	HW	224	99	1	0
		KB	68,306	99	1	0

5.2.3.5 Grades 9–12

Table 5.2.3.5.1

Interrater Agreement: Writ 9-12 A S502 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	94	96	4	0
		KB	27,528	97	3	0
	2	HW	94	100	0	0
		KB	27,506	97	3	0

Table 5.2.3.5.2

Interrater Agreement: Writ 9-12 B/C S502 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	64	100	0	0
		KB	59,702	98	2	0
	2	HW	64	100	0	0
		KB	61,148	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 S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	4,674	98	2	0
	2	4,442	98	2	0
	3	4,678	98	2	0

Table 5.2.4.1.2

Interrater Agreement: Spek 1 A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	32,794	99	1	0
	2	32,792	86	14	0
	3	32,578	99	1	0
	4	32,574	89	11	0
	5	33,020	99	1	0
	6	33,020	89	11	0

Table 5.2.4.1.3

Interrater Agreement: Spek 1 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	35,088	82	18	0
	2	35,082	84	16	0
	3	34,640	86	14	0
	4	34,640	79	21	0
	5	35,892	84	16	0
	6	35,892	81	19	0

5.2.4.2 Grade 2–3

Table 5.2.4.2.1

Interrater Agreement: Spek 2-3 Pre-A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	6,322	99	1	0
	2	6,270	99	1	0
	3	5,640	99	1	0

Table 5.2.4.2.2

Interrater Agreement: Spek 2-3 A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	39,900	99	1	0
	2	39,890	84	15	0
	3	40,016	99	1	0
	4	40,016	83	17	0
	5	40,796	100	0	0
	6	40,796	83	16	1

Table 5.2.4.2.3

Interrater Agreement: Spek 2-3 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	94,252	82	17	0
	2	94,252	81	19	0
	3	91,846	81	19	0
	4	91,846	78	21	0
	5	94,366	82	17	0
	6	94,356	78	21	1

5.2.4.3 Grades 4–5

Table 5.2.4.3.1

Interrater Agreement: Spek 4-5 Pre-A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	2,792	97	3	0
	2	2,514	98	2	0
	3	2,256	98	2	0

Table 5.2.4.3.2

Interrater Agreement: Spek 4-5 A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	17,730	98	2	0
	2	17,730	87	13	0
	3	17,086	99	1	0
	4	17,086	90	10	0
	5	16,652	100	0	0
	6	16,654	87	13	0

Table 5.2.4.3.3

Interrater Agreement: Spek 4-5 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	98,476	85	15	0
	2	98,476	82	18	0
	3	100,424	84	16	0
	4	100,426	82	18	0
	5	99,212	84	16	0
	6	99,212	80	20	0

5.2.4.4 Grades 6–8

Table 5.2.4.4.1

Interrater Agreement: Spek 6-8 Pre-A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	4,248	99	1	0
	2	2,928	98	2	0
	3	3,246	98	2	0

Table 5.2.4.4.2

Interrater Agreement: Spek 6-8 A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	25,914	100	0	0
	2	25,914	91	9	0
	3	25,958	99	1	0
	4	25,958	85	14	1
	5	25,912	99	1	0
	6	25,912	90	10	0

Table 5.2.4.4.3

Interrater Agreement: Spek 6-8 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	96,148	86	13	0
	2	96,150	85	15	0
	3	95,350	81	19	1
	4	95,350	81	18	0
	5	93,116	85	15	0
	6	93,116	80	20	0

5.2.4.5 Grades 9–12

Table 5.2.4.5.1

Interrater Agreement: Spek 9-12 Pre-A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	7,804	99	1	0
	2	7,750	99	1	0
	3	7,482	99	1	0

Table 5.2.4.5.2

Interrater Agreement: Spek 9-12 A S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	40,084	100	0	0
	2	40,084	84	16	0
	3	39,368	100	0	0
	4	39,368	85	15	0
	5	39,458	100	0	0
	6	39,458	86	14	0

Table 5.2.4.5.3

Interrater Agreement: Spek 9-12 B/C S502 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	58,348	77	22	0
	2	58,348	79	20	0
	3	57,402	79	21	0
	4	57,402	80	19	0
	5	60,230	81	19	0
	6	60,230	81	19	0

5.3 Conditional Standard Errors of Measurement of the Scale Scores at the Cut Points

The tables in this section present information about the CSEM 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 score at that point on the underlying latent ability scale. We first computed the CSEM on the theta metric, which is the square root of the inverse of the Test Information Function, and then linearly transformed the values to the ACCESS scale score metric using the multiplicative constant of the linear equation for the domain (See Section 2.2). The CSEM value based on IRT can vary across test scores. For example, in the Listening and Reading domain, if a student gets either a very few or a very large number of items correct (i.e., scores at the extremes of the score distribution), the CSEM will be greater in value than it would be if the student gets a moderate number of items correct. Scores near the middle of the score distribution typically have lower CSEM compared to the extremes because many tests are comprised of a large proportion of moderately difficult items which are suited to measuring students of moderate proficiency. The CSEM can be used to construct the error band quantifying uncertainty in a student’s score. An approximate 68% confidence interval for a scale score is given by one CSEM below the scale score and one CSEM above the scale score. 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 WIDA states vary in how decisions about the ACCESS scores are made. As discussed in Section 5, all WIDA states use composite scale scores in making reclassification decision and no WIDA state uses a single domain scale score in making reclassification decision. Because the cut points depend on the grade level, we provide information for each grade level within a grade-level cluster.

Since ACCESS test scores were scaled using the IRT method, CSEM values for the scale scores at the highest cut points are typically high. The IRT method is known to produce higher CSEMs 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, the numbers of students at the highest cut points are typically smaller than those at other cut points. Therefore, the measurement errors associated with the scale scores at the highest cut points tend to be higher than those of the scale scores at the lower cut points since there are fewer students available for estimating the scores and the measurement errors of these scores.

Since the Listening and Reading tests are multistage adaptive tests, the CSEM 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 for all students' scale scores at the cut point by grade level. Note that there are some rare cases where there are no observed scale scores corresponding to the cut points; therefore, we cannot provide these descriptive statistics. Because Listening and Reading tests are multistage adaptive tests, we would not expect a large variation in the mean CSEM of students' scale score across cut points within a grade level.

For Writing and Speaking, we present the CSEMs for the scale scores by tier. From these tables, it is possible to identify how well the different Writing and Speaking tiers are targeted for making decisions about students at the various proficiency level cut points. For example, Tier A is intended for students at the lower end of the language proficiency continuum. Therefore, the CSEMs of the Tier A student scale scores are expected to be lower at the 1/2 and at the 2/3 proficiency level cut points as compared to those at the 4/5 cut. These tables provide comparable information on how well the two-tier forms are targeted to provide the most accurate measure to place their intended students into the language proficiency levels that they target. 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 is given for scale scores at each cut point.

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	1	236	364	18.88	18.88	18.88	0.00
2/3	1	259	38	16.33	16.33	16.33	0.00
3/4	1	291	163	16.33	17.86	16.34	0.13
4/5	1	303	1,090	16.84	17.35	17.28	0.18
5/6	1	327	104	17.86	19.39	19.12	0.58

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	2	245	1,485	20.41	20.41	20.41	0.00
	3	262	1	17.86	17.86	17.86	0.00
2/3	2	283	669	17.35	18.88	17.37	0.19
	3	300	1,042	17.86	18.88	18.06	0.32
3/4	2	314	2,743	18.88	19.90	19.37	0.14
	3	331	1,439	18.37	19.90	18.76	0.59
4/5	2	330	489	18.88	19.90	19.30	0.33
	3	349	346	19.90	22.45	21.22	0.90
5/6	2	354	70	21.43	23.47	21.65	0.54
	3	374	N/A	N/A	N/A	N/A	N/A

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	4	275	4	17.86	20.41	19.77	1.28
	5	285	7	17.86	17.86	17.86	0.00
2/3	4	313	8	16.84	16.84	16.84	0.00
	5	323	N/A	N/A	N/A	N/A	N/A
3/4	4	343	183	17.35	18.88	18.85	0.17
	5	354	N/A	N/A	N/A	N/A	N/A
4/5	4	363	239	17.86	18.88	18.61	0.30
	5	375	161	18.37	19.39	18.58	0.37
5/6	4	388	112	18.37	18.88	18.87	0.05
	5	401	2	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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	6	294	40	20.92	20.92	20.92	0.00
	7	302	172	21.43	21.43	21.43	0.00
	8	308	88	18.37	20.41	19.94	0.86
2/3	6	332	8	16.33	16.33	16.33	0.00
	7	340	21	16.33	16.33	16.33	0.00
	8	347	109	16.84	18.88	16.87	0.28
3/4	6	363	32	15.82	16.33	16.29	0.13
	7	370	8	15.82	17.35	16.39	0.51
	8	377	601	16.84	17.35	16.89	0.16
4/5	6	385	1,892	16.84	17.35	17.22	0.22
	7	394	334	16.84	17.86	17.24	0.28
	8	402	121	17.35	18.37	17.91	0.31
5/6	6	411	40	17.35	17.86	17.81	0.16
	7	420	88	17.86	18.37	18.18	0.25
	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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	9	314	1,434	18.88	20.92	19.07	0.40
	10	325	N/A	N/A	N/A	N/A	N/A
	11	335	N/A	N/A	N/A	N/A	N/A
	12	342	7	17.35	18.88	18.44	0.75
2/3	9	353	28	16.33	16.33	16.33	0.00
	10	358	1,406	16.33	16.33	16.33	0.00
	11	364	29	16.33	16.33	16.33	0.00
	12	368	41	16.33	16.84	16.46	0.23
3/4	9	383	42	16.84	17.35	17.04	0.25
	10	389	1,260	16.84	16.84	16.84	0.00
	11	394	216	17.35	17.35	17.35	0.00
	12	398	29	17.35	17.35	17.35	0.00
4/5	9	409	177	17.35	17.86	17.59	0.26
	10	415	343	17.86	19.39	18.37	0.17
	11	420	327	17.86	18.37	18.17	0.25
	12	426	231	18.37	21.43	19.08	1.24
5/6	9	434	2,950	17.86	19.90	17.87	0.14
	10	441	266	18.37	20.92	19.11	1.15
	11	447	145	20.41	22.45	20.77	0.79
	12	452	N/A	N/A	N/A	N/A	N/A

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	1	264	1,365	10.71	12.76	11.80	0.53
2/3	1	286	1,070	9.69	10.71	9.96	0.28
3/4	1	304	1,164	9.69	10.20	10.17	0.12
4/5	1	315	64	9.69	10.20	10.12	0.19
5/6	1	334	1,804	10.20	10.71	10.20	0.02

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	2	283	277	10.71	12.76	12.03	0.41
	3	297	1,683	9.69	10.71	10.71	0.05
2/3	2	307	2,520	9.69	10.20	10.20	0.05
	3	323	634	9.69	10.20	9.73	0.13
3/4	2	326	632	9.69	10.20	10.10	0.20
	3	342	208	9.69	10.20	10.03	0.24
4/5	2	337	5,335	9.69	10.20	9.70	0.04
	3	352	394	10.20	10.71	10.34	0.23
5/6	2	355	N/A	N/A	N/A	N/A	N/A
	3	370	102	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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	4	307	505	10.20	12.24	11.45	0.37
	5	316	277	9.69	11.73	11.17	0.51
2/3	4	335	1,181	9.69	11.22	10.10	0.62
	5	345	45	9.69	10.71	9.99	0.28
3/4	4	354	402	9.69	10.71	10.32	0.30
	5	364	361	10.20	10.71	10.21	0.04
4/5	4	364	757	10.20	10.20	10.20	0.00
	5	373	485	10.20	10.71	10.25	0.15
5/6	4	382	5	10.20	11.22	10.92	0.46
	5	391	44	11.22	11.22	11.22	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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	6	323	220	11.73	12.24	12.00	0.26
	7	329	141	11.73	12.24	11.83	0.20
	8	335	122	11.73	11.73	11.73	0.00
2/3	6	353	3,709	10.20	10.71	10.22	0.08
	7	360	345	10.20	10.71	10.24	0.12
	8	366	733	9.69	11.22	10.21	0.08
3/4	6	373	175	9.69	10.71	10.34	0.23
	7	380	572	10.20	10.71	10.51	0.25
	8	386	141	10.20	12.24	10.56	0.41
4/5	6	382	292	10.20	11.73	10.56	0.25
	7	389	1,213	10.71	11.22	10.73	0.09
	8	395	274	10.71	11.73	10.97	0.31
5/6	6	399	3	11.22	11.22	11.22	0.00
	7	406	40	11.22	11.73	11.49	0.26
	8	412	109	11.73	12.24	11.79	0.15

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	9	340	757	11.73	12.24	11.74	0.05
	10	344	37	11.73	12.76	11.86	0.28
	11	348	184	11.22	12.24	11.44	0.34
	12	352	5	11.22	11.73	11.33	0.23
2/3	9	372	97	10.20	10.20	10.20	0.00
	10	377	378	9.69	10.20	9.91	0.25
	11	382	397	9.69	10.71	9.82	0.27
	12	386	267	9.69	10.71	9.77	0.24
3/4	9	392	397	9.69	10.71	10.12	0.19
	10	397	382	9.69	11.22	10.03	0.26
	11	402	90	9.69	10.71	10.20	0.11
	12	407	858	10.20	10.71	10.21	0.02
4/5	9	401	203	9.69	10.20	10.19	0.09
	10	406	156	9.69	10.71	10.47	0.26
	11	410	77	10.20	10.71	10.28	0.18
	12	414	41	10.20	11.22	10.66	0.34
5/6	9	418	31	10.20	11.22	10.39	0.39
	10	423	30	10.71	11.22	11.17	0.16
	11	427	55	10.71	11.22	11.18	0.15
	12	432	3	11.22	11.22	11.22	0.00

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	1	238	14.77	14.50
2/3	1	275	20.68	18.80
3/4	1	337	20.94	21.75
4/5	1	382	18.80	18.80
5/6	1	405	23.63	20.14

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	2	242	14.77	14.50
	3	247	15.57	15.04
2/3	2	279	20.68	20.14
	3	283	20.94	20.68
3/4	2	341	20.94	21.21
	3	346	20.41	20.79
4/5	2	388	19.33	18.80
	3	394	19.87	19.33
5/6	2	411	24.43	23.09
	3	418	27.12	25.51

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	4	266	14.23	28.46
	5	267	14.23	27.66
2/3	4	288	17.18	16.72
	5	293	17.99	15.31
3/4	4	351	21.75	20.68
	5	356	21.75	20.94
4/5	4	401	18.80	21.48
	5	407	18.53	21.21
5/6	4	425	19.60	19.87
	5	433	21.21	19.33

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	6	268	14.50	14.77
	7	273	15.31	14.23
	8	281	16.65	14.23
2/3	6	298	19.60	16.92
	7	305	20.41	18.26
	8	311	20.94	19.33
3/4	6	361	21.48	22.02
	7	367	20.94	21.75
	8	372	20.68	21.48
4/5	6	413	18.80	18.80
	7	419	19.33	18.53
	8	424	20.41	18.53
5/6	6	441	25.24	20.46
	7	450	29.27	22.82
	8	459	34.64	26.31

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	9	289	14.50	14.77
	10	298	14.77	14.95
	11	308	16.11	16.11
	12	318	17.82	17.78
2/3	9	319	17.99	17.99
	10	326	19.33	19.33
	11	335	20.41	20.41
	12	344	21.21	21.21
3/4	9	378	21.75	21.75
	10	385	21.48	21.48
	11	391	21.21	21.21
	12	398	20.94	20.75
4/5	9	430	18.80	18.80
	10	436	18.80	18.80
	11	441	18.95	19.01
	12	447	19.33	19.60
5/6	9	469	24.43	24.70
	10	479	28.73	29.00
	11	490	34.91	35.18
	12	501	42.69	43.23

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	1	205	21.35	15.50
2/3	1	261	28.08	19.89
3/4	1	311	23.98	17.26
4/5	1	361	31.00	21.06
5/6	1	403	53.52	34.80

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	2	220	23.98	16.67
	3	234	26.62	17.84
2/3	2	273	26.91	19.60
	3	283	26.03	19.01
3/4	2	322	24.28	17.26
	3	332	24.86	17.26
4/5	2	374	35.68	22.52
	3	386	41.53	26.03
5/6	2	415	63.76	38.31
	3	425	75.17	44.75

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	4	246	22.81	15.79
	5	258	24.86	16.67
2/3	4	293	28.08	19.30
	5	302	27.49	19.60
3/4	4	342	23.98	17.84
	5	350	23.98	17.55
4/5	4	397	31.29	19.60
	5	407	34.80	21.35
5/6	4	435	51.18	29.25
	5	443	57.62	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.

5.3.4.4 Grades 6–8

Table 5.3.4.4

Conditional Standard Errors of Measurement of Scale Scores at the Cut Points: Spek 6-8 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	6	268	23.11	15.79
	7	277	24.57	16.67
	8	284	25.74	17.26
2/3	6	310	28.37	19.89
	7	317	28.37	19.89
	8	323	27.79	19.89
3/4	6	360	24.28	17.55
	7	369	23.98	17.26
	8	377	23.98	16.96
4/5	6	417	30.71	19.89
	7	425	33.63	21.35
	8	433	36.85	23.40
5/6	6	451	47.09	29.25
	7	457	51.48	31.59
	8	463	56.45	34.51

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 S502 Online

Proficiency Level Cut Point	Grade	Cut Score	CSEM	
			Tier A	Tier B/C
1/2	9	290	24.86	16.96
	10	295	25.74	17.26
	11	299	26.32	17.84
	12	302	26.91	18.13
2/3	9	328	28.37	19.89
	10	333	28.08	19.89
	11	337	27.79	19.89
	12	340	27.49	19.60
3/4	9	385	23.98	17.14
	10	393	24.28	17.26
	11	400	24.86	17.26
	12	406	25.45	17.55
4/5	9	440	34.51	22.23
	10	446	37.14	23.69
	11	451	39.78	24.86
	12	455	41.82	26.32
5/6	9	468	50.60	31.00
	10	471	52.94	32.17
	11	474	55.28	33.63
	12	476	57.03	34.51

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 score or scaled scores would agree with those made based on their true scores (Livingston, 2018; Livingston & Lewis, 1995). True scores are assumed to be measured perfectly but are unknown. Therefore, to provide the best estimation of classification accuracy, we use test data from one test administration to estimate the true scores based on observed scores and the parameters of the model used in estimating the true scores. It is then possible 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 was implemented here, uses information about the reliability of the students' test scores, the cut points, and the observed distribution of scores. Then, using a four-parameter beta distribution, we model the distribution of the true scores and of 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 who would be classified into the same language proficiency level by both their observed test scores and their true scores. For example, an overall classification accuracy index of 0.774 means that 77% of the students would be classified into the correct proficiency level across all six proficiency levels according to their observed and true scores. **Overall classification consistency** indicates the percentage of all students who would be classified into the same language proficiency level by both the administered test and by a parallel test. For example, an overall classification consistency index of 0.664 means that 66% of the students would be classified into the same proficiency level if two parallel forms were administered. Classification consistency indices are always lower than the corresponding classification accuracy indices, because in classification consistency, classifications based on students' performance on the administered test and classification based on students' performance on a parallel test are both subject to measurement error. In contrast, in classification accuracy, only the classification based on students' performance on the administered test contains error while classification based on students' true score is assumed to be free of measurement error.

Marginal Classification Accuracy and Consistency

Overall classification accuracy and consistency indices indicate the degree to which students are accurately and consistently classified in the same WIDA proficiency levels, but not the degree to which students are accurately or consistently classified 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 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 scale scores at the cut points** report the percentage of students who are accurately placed into proficiency levels above and below each cut point based on their scale scores. For example, a classification accuracy index of 0.774 at the PL 4/PL 5 cut point means that 77% of the students would be classified in the same way if they were classified according to their observed scale score and their true scale score, 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 scale scores at the cut points** report the percentage of students classified consistently above and below each cut point based on their scale scores. For example, a classification consistency index of 0.664 at the PL 4/PL 5 cut point means that 66% of the students would be classified in the same way if two parallel forms were administered, 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' scale scores at the cut points than the overall classification accuracy and consistency (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 calculation of classification accuracy and consistency indices is affected by the interactions of a number of factors: (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 a 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' 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 the test score reliability coefficient for ACCESS Online Writing domain raw scores is estimated based on only two tasks. Therefore, the classification accuracy and consistency indices for the Writing domain might be lower than those of other domains as a result.

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 scale scores at the cut points for each grade level. The third reports the marginal classification consistency indices based on 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 fewer than 200 students were classified 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 very 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 0.584 to 0.805, and the overall classification consistency indices ranged from 0.483 to 0.751. The lowest overall classification accuracy and consistency indices for Listening were at Grade 11.

For Reading, as shown in Table 5.4.2.1, the overall classification accuracy indices ranged from 0.614 to 0.693, and the overall classification consistency indices ranged from 0.507 to 0.598. The lowest overall classification accuracy index for Reading was found for students in Grade 1, while the lowest overall classification consistency was found for students in Grade 2 and Grade 3.

For Writing, as shown in Table 5.4.3.1, the overall classification accuracy indices ranged from 0.567 to 0.786, and the overall classification consistency indices ranged from 0.501 to 0.699. The lowest overall classification accuracy and consistency indices for Writing was at Grade 5.

For Speaking, as shown in Table 5.4.4.1, the overall classification accuracy indices ranged from 0.657 to 0.733, and the overall classification consistency indices ranged from 0.550 to 0.640. The lowest overall classification accuracy index for Speaking was at Grade 5 and 7, while the lowest overall classification consistency was at Grade 5.

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 the range of the marginal classification accuracy and consistency indices for the domains across grades and highlight the grades with the lowest marginal classification accuracy and the lowest classification consistency by domain.

For Listening, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.866 to 0.990 (Table 5.4.1.2), and the marginal classification consistency indices ranged from 0.815 to 0.988 (Table 5.4.1.3). The lowest overall marginal classification accuracy and consistency indices were at the PL 5/6 cut point of Grade 6.

For Reading, the marginal classification accuracy indices for scale scores at the cut points ranged from 0.864 to 0.979 (Table 5.4.2.2), and the marginal classification consistency indices ranged from 0.813 to 0.969 (Table 5.4.2.3). Grade 1, at the PL 1/PL 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/PL 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 0.80-0.90 range.

For Writing, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.675 to 0.997 (Table 5.4.3.2), and the marginal classification consistency indices ranged from 0.613 to 0.997 (Table 5.4.3.3). Grade 5 at the PL 3/PL 4 cut point had the lowest marginal classification accuracy index and Grade 4 at the PL 3/4 cut point had the lowest

consistency index. Note that Grade 5 also had the lowest overall classification accuracy and consistency indices in the Writing domain followed by Grade 4. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut point appeared to have contributed to its low overall classification accuracy and consistency.

For Speaking, the marginal classification accuracy indices based on scale scores at the cut points ranged from 0.829 to 0.997 (Table 5.4.4.2), and the marginal classification consistency indices ranged from 0.773 to 0.997 (Table 5.4.4.3). Grade 9 at the PL 2/PL 3 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 9 Speaking are still in the 0.70-0.90 range.

The overall and marginal classification accuracy and consistency indices produced similar results, i.e., when overall classification accuracy and consistency indices were low, marginal classification accuracy and consistency indices tended to be low.

On average, we observed that the marginal classification and consistency indices for PL cut scores in the middle (e.g., the PL 2/PL 3 and PL 3/PL 4 cut points) tended to have low marginal classification accuracy and consistency indices across the domains of Reading, Writing, and Speaking. This finding is consistent with findings from previous researchers (Ercikan & Julian, 2002; Lee et al., 2002), who have reported that marginal classification accuracy and consistency at cut points in the middle of the proficiency level range tend to be lower than for cut points in the lower and upper ends. 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. (Cut points tend to be closer to each other when the number of proficiency levels is high.) Marginal classification accuracy and consistency are expected to vary for different ability levels due to variation in measurement accuracy. The further away the students' 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 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 students would be misclassified (Ercikan & Julian, 2002). Since ACCESS has six proficiency levels, the intervals between cut scores in the middle occupy relatively narrow ranges on the ability scale as compared to other proficiency levels. The marginal classification accuracy and consistency based on the scale scores for the 2/3 and 3/4 cut points are lower than for other cut points, as might be expected.

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 S502 Online

Grade	Accuracy	Consistency
1	0.662	0.590
2	0.598	0.511
3	0.602	0.517
4	0.805	0.751
5	0.736	0.672
6	0.640	0.542
7	0.617	0.523
8	0.600	0.508
9	0.604	0.504
10	0.589	0.491
11	0.584	0.483
12	0.588	0.490

Table 5.4.1.2

Classification Accuracy Indices at Cut Score Level: List S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.950	0.935	0.903	0.900	0.891
2	0.944	0.901	0.884	0.896	0.898
3	0.950	0.913	0.894	0.889	0.884
4	0.990	0.980	0.957	0.945	0.904
5	0.986	0.972	0.938	0.920	0.877
6	0.990	0.965	0.912	0.885	0.866
7	0.986	0.949	0.887	0.881	0.881
8	0.977	0.941	0.886	0.875	0.879
9	0.966	0.919	0.883	0.898	0.911
10	0.949	0.920	0.879	0.894	0.915
11	0.947	0.923	0.882	0.883	0.916
12	0.938	0.909	0.885	0.902	0.920

Table 5.4.1.3

Classification Consistency Indices at Cut Score Level: List S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.932	0.906	0.865	0.858	0.846
2	0.920	0.862	0.839	0.848	0.858
3	0.928	0.879	0.850	0.842	0.840
4	0.988	0.971	0.940	0.916	0.861
5	0.981	0.957	0.913	0.883	0.830
6	0.986	0.947	0.877	0.835	0.815
7	0.979	0.923	0.849	0.830	0.833
8	0.968	0.912	0.845	0.823	0.831
9	0.951	0.884	0.841	0.853	0.876
10	0.930	0.882	0.836	0.849	0.881
11	0.927	0.886	0.838	0.838	0.880
12	0.913	0.871	0.841	0.859	0.889

5.4.2 Reading

Table 5.4.2.1

Overall Accuracy and Consistency of Classification Indices: Read S502 Online

Grade	Accuracy	Consistency
1	0.614	0.508
2	0.620	0.507
3	0.616	0.507
4	0.615	0.514
5	0.629	0.529
6	0.693	0.598
7	0.680	0.585
8	0.678	0.586
9	0.664	0.566
10	0.664	0.569
11	0.666	0.571
12	0.668	0.572

Table 5.4.2.2

Classification Accuracy Indices at Cut Score Level: Read S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.864	0.895	0.920	0.941	0.967
2	0.945	0.886	0.895	0.913	0.960
3	0.908	0.884	0.915	0.922	0.953
4	0.950	0.918	0.892	0.889	0.933
5	0.939	0.905	0.900	0.905	0.942
6	0.918	0.896	0.929	0.950	0.979
7	0.915	0.899	0.929	0.942	0.965
8	0.910	0.903	0.931	0.941	0.959
9	0.943	0.908	0.916	0.921	0.943
10	0.936	0.912	0.918	0.921	0.943
11	0.938	0.914	0.919	0.917	0.940
12	0.937	0.915	0.915	0.915	0.946

Table 5.4.2.3

Classification Consistency Indices at Cut Score Level: Read S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.813	0.851	0.889	0.916	0.953
2	0.920	0.841	0.851	0.878	0.941
3	0.869	0.837	0.876	0.892	0.933
4	0.930	0.882	0.852	0.850	0.903
5	0.914	0.866	0.860	0.869	0.916
6	0.886	0.858	0.897	0.927	0.969
7	0.881	0.862	0.897	0.916	0.950
8	0.874	0.867	0.899	0.915	0.943
9	0.920	0.871	0.879	0.889	0.920
10	0.911	0.876	0.883	0.890	0.920
11	0.913	0.878	0.884	0.886	0.915
12	0.912	0.879	0.882	0.885	0.923

5.4.3 Writing

Table 5.4.3.1

Overall Accuracy and Consistency of Classification Indices: Writ S502 Online

Grade	Accuracy	Consistency
1	0.707	0.598
2	0.724	0.646
3	0.664	0.608
4	0.664	0.534
5	0.567	0.501
6	0.786	0.699
7	0.695	0.615
8	0.755	0.641
9	0.666	0.563
10	0.705	0.588
11	0.671	0.567
12	0.669	0.565

Table 5.4.3.2

Classification Accuracy Indices at Cut Score Level: Writ S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.855	0.852	0.993	N/A	N/A
2	0.950	0.862	0.910	N/A	N/A
3	0.975	0.940	0.749	N/A	N/A
4	0.980	0.961	0.718	0.982	0.995
5	0.974	0.946	0.675	0.962	0.997
6	0.963	0.893	0.929	N/A	N/A
7	0.955	0.880	0.858	N/A	N/A
8	0.948	0.888	0.915	N/A	N/A
9	0.953	0.876	0.836	0.996	N/A
10	0.945	0.865	0.894	0.994	N/A
11	0.922	0.853	0.897	0.992	N/A
12	0.929	0.848	0.885	N/A	N/A

Table 5.4.3.3

Classification Consistency Indices at Cut Score Level: Writ S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.799	0.764	0.993	N/A	N/A
2	0.927	0.819	0.886	N/A	N/A
3	0.963	0.920	0.720	N/A	N/A
4	0.970	0.942	0.613	0.975	0.995
5	0.961	0.927	0.620	0.950	0.997
6	0.942	0.850	0.897	N/A	N/A
7	0.931	0.840	0.833	N/A	N/A
8	0.923	0.841	0.857	N/A	N/A
9	0.927	0.833	0.786	0.994	N/A
10	0.914	0.811	0.842	0.992	N/A
11	0.888	0.799	0.855	0.990	N/A
12	0.891	0.791	0.850	N/A	N/A

5.4.4 Speaking

Table 5.4.4.1

Overall Accuracy and Consistency of Classification Indices: Spek S502 Online

Grade	Accuracy	Consistency
1	0.666	0.562
2	0.676	0.566
3	0.672	0.553
4	0.672	0.562
5	0.657	0.550
6	0.662	0.573
7	0.657	0.562
8	0.705	0.597
9	0.667	0.576
10	0.721	0.629
11	0.733	0.638
12	0.729	0.640

Table 5.4.4.2

Classification Accuracy Indices at Cut Score Level: Spek S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.921	0.856	0.891	0.994	N/A
2	0.923	0.851	0.903	0.992	N/A
3	0.962	0.987	0.871	0.991	0.997
4	0.948	0.880	0.860	0.982	N/A
5	0.933	0.860	0.868	0.992	N/A
6	0.936	0.869	0.856	N/A	N/A
7	0.925	0.851	0.878	0.996	N/A
8	0.925	0.867	0.908	N/A	N/A
9	0.904	0.829	0.929	N/A	N/A
10	0.902	0.855	0.956	N/A	N/A
11	0.912	0.848	0.966	N/A	N/A
12	0.907	0.838	0.978	N/A	N/A

Table 5.4.4.3

Classification Consistency Indices at Cut Score Level: Spek S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.883	0.800	0.859	0.992	N/A
2	0.887	0.794	0.866	0.991	N/A
3	0.904	0.808	0.814	0.991	0.997
4	0.923	0.833	0.805	0.980	N/A
5	0.901	0.809	0.822	0.991	N/A
6	0.906	0.820	0.829	N/A	N/A
7	0.891	0.801	0.846	0.996	N/A
8	0.893	0.812	0.862	N/A	N/A
9	0.862	0.773	0.908	N/A	N/A
10	0.862	0.795	0.935	N/A	N/A
11	0.874	0.786	0.948	N/A	N/A
12	0.868	0.777	0.966	N/A	N/A

5.5 Reliabilities of Students' Composite 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 can be explained by the variance in their true composite scale scores.

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 .

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 derive a single reliability coefficient for the grade-level cluster. To produce this single value, we weight 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 0.88 to 0.91. The stratified Cronbach's coefficient alphas for the Oral scale scores ranged from 0.89 to 0.91 for male students; 0.88 to 0.90 for female students; 0.88 to 0.91 for Hispanic students; 0.87 to 0.89 for non-Hispanic students; and 0.87 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 0.90. The stratified Cronbach's coefficient alphas for the Literacy scale scores ranged from 0.88 to 0.91 for male students; 0.86 to 0.89 for female students; 0.86 to

0.90 for Hispanic students; 0.87 to 0.90 for non-Hispanic students; and 0.84 to 0.90 for students with an IEP.

The stratified Cronbach's coefficient alphas for the Comprehension scale scores computed for all students ranged from 0.91 to 0.94. The stratified Cronbach's coefficient alphas for the Comprehension scale scores ranged from 0.91 to 0.94 for male students; 0.91 to 0.93 for female students; 0.89 to 0.93 for Hispanic students; 0.91 to 0.93 for non-Hispanic students; and 0.89 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 0.92 to 0.94. The stratified Cronbach's coefficient alphas for the Overall scale scores ranged from 0.92 to 0.94 for male students; 0.91 to 0.93 for female students; 0.91 to 0.94 for Hispanic students; 0.92 to 0.94 for non-Hispanic students; and 0.91 to 0.94 for students with an IEP.

5.5.1 Oral

Table 5.5.1.1

Reliabilities of Composite Scale Scores: Oral S502 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.50	3035.78	0.87
	Speaking	0.50	2789.67	0.82
	Oral		2242.06	0.90
2-3	Listening	0.50	3523.26	0.86
	Speaking	0.50	2839.14	0.82
	Oral		2504.34	0.90
4-5	Listening	0.50	2400.06	0.82
	Speaking	0.50	2416.71	0.83
	Oral		1840.55	0.88
6-8	Listening	0.50	2116.22	0.84
	Speaking	0.50	2639.83	0.84
	Oral		1837.46	0.90
9-12	Listening	0.50	2731.05	0.87
	Speaking	0.50	2978.28	0.83
	Oral		2282.14	0.91

Table 5.5.1.2

Reliabilities of Composite Scale Scores: Oral S502 Online by Gender

Cluster	Component	Weight	Female		Male	
			Variance	Reliability	Variance	Reliability
1	Listening	0.50	2895.55	0.86	3123.85	0.87
	Speaking	0.50	2783.12	0.82	2700.03	0.82
	Oral		2168.11	0.89	2244.23	0.90
2-3	Listening	0.50	3306.82	0.85	3688.97	0.86
	Speaking	0.50	2779.17	0.82	2799.85	0.82
	Oral		2374.82	0.89	2568.86	0.90
4-5	Listening	0.50	2220.31	0.81	2527.81	0.82
	Speaking	0.50	2363.33	0.82	2427.28	0.83
	Oral		1733.36	0.88	1909.39	0.89
6-8	Listening	0.50	2003.33	0.84	2192.58	0.85
	Speaking	0.50	2670.55	0.84	2597.14	0.84
	Oral		1796.25	0.90	1857.95	0.90
9-12	Listening	0.50	2630.65	0.86	2791.66	0.87
	Speaking	0.50	2940.86	0.83	2988.70	0.84
	Oral		2239.73	0.90	2296.10	0.91

Table 5.5.1.3

Reliabilities of Composite Scale Scores: Oral S502 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.50	2903.93	0.87	3122.55	0.86
	Speaking	0.50	2730.26	0.82	2682.23	0.81
	Oral		2148.09	0.90	2225.47	0.89
2-3	Listening	0.50	3443.16	0.86	3469.19	0.84
	Speaking	0.50	2872.60	0.82	2618.21	0.81
	Oral		2486.32	0.90	2358.13	0.89
4-5	Listening	0.50	2342.82	0.82	2256.92	0.80
	Speaking	0.50	2396.52	0.83	2200.27	0.81
	Oral		1800.61	0.88	1671.29	0.87
6-8	Listening	0.50	2054.43	0.84	2087.35	0.83
	Speaking	0.50	2636.49	0.84	2315.36	0.82
	Oral		1794.62	0.90	1696.88	0.89
9-12	Listening	0.50	2702.41	0.87	2550.36	0.85
	Speaking	0.50	3001.12	0.84	2598.11	0.81
	Oral		2269.73	0.91	2015.23	0.89

Table 5.5.1.4

Reliabilities of Composite Scale Scores: Oral S502 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.50	3203.66	0.89
	Speaking	0.50	3037.48	0.83
	Oral		2407.18	0.91
2-3	Listening	0.50	3591.65	0.88
	Speaking	0.50	3070.38	0.81
	Oral		2663.94	0.90
4-5	Listening	0.50	2183.79	0.82
	Speaking	0.50	2242.96	0.81
	Oral		1610.79	0.87
6-8	Listening	0.50	1691.47	0.81
	Speaking	0.50	2263.24	0.83
	Oral		1414.28	0.88
9-12	Listening	0.50	1888.93	0.82
	Speaking	0.50	2530.12	0.83
	Oral		1607.15	0.88

5.5.2 Literacy

Table 5.5.2.1

Reliabilities of Composite Scale Scores: Litr S502 Online

Cluster	Component	Weight	Variance	Reliability
1	Reading	0.50	1115.27	0.88
	Writing	0.50	1716.10	0.78
	Literacy			1052.54
2-3	Reading	0.50	940.55	0.87
	Writing	0.50	2166.15	0.84
	Literacy			1161.19
4-5	Reading	0.50	1232.06	0.90
	Writing	0.50	2115.64	0.74
	Literacy			1344.27
6-8	Reading	0.50	1381.32	0.90
	Writing	0.50	1369.06	0.77
	Literacy			1115.03
9-12	Reading	0.50	1339.30	0.91
	Writing	0.50	1416.15	0.74
	Literacy			1055.21

Table 5.5.2.2

Reliabilities of Composite Scale Scores: Litr S502 Online by Gender

Cluster	Component	Weight	Female		Male	
			Variance	Reliability	Variance	Reliability
1	Reading	0.50	1157.01	0.89	1070.48	0.88
	Writing	0.50	1544.24	0.77	1829.14	0.79
	Literacy			1020.29	0.88	1061.72
2-3	Reading	0.50	931.51	0.87	943.47	0.87
	Writing	0.50	1989.64	0.82	2243.42	0.85
	Literacy			1106.22	0.89	1178.74
4-5	Reading	0.50	1153.31	0.89	1284.99	0.90
	Writing	0.50	1846.90	0.71	2264.83	0.76
	Literacy			1214.22	0.86	1415.88
6-8	Reading	0.50	1335.66	0.90	1403.38	0.91
	Writing	0.50	1341.14	0.75	1344.59	0.78
	Literacy			1087.64	0.89	1109.71
9-12	Reading	0.50	1293.46	0.90	1360.44	0.91
	Writing	0.50	1393.59	0.73	1389.63	0.75
	Literacy			1032.83	0.88	1044.74

Table 5.5.2.3

Reliabilities of Composite Scale Scores: Litr S502 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Reading	0.50	858.77	0.85	1413.82	0.90
	Writing	0.50	1605.82	0.79	1706.31	0.76
	Literacy			853.61	0.86	1231.65
2-3	Reading	0.50	851.72	0.86	1036.37	0.88
	Writing	0.50	2288.70	0.85	1719.11	0.81
	Literacy			1145.42	0.90	1063.60
4-5	Reading	0.50	1169.90	0.89	1291.10	0.90
	Writing	0.50	2127.81	0.74	1792.12	0.72
	Literacy			1313.83	0.87	1252.95
6-8	Reading	0.50	1298.25	0.90	1503.52	0.91
	Writing	0.50	1354.40	0.78	1239.60	0.76
	Literacy			1070.06	0.90	1111.76
9-12	Reading	0.50	1268.08	0.90	1393.66	0.91
	Writing	0.50	1405.12	0.74	1285.76	0.73
	Literacy			1018.37	0.88	1012.95

Table 5.5.2.4

Reliabilities of Composite Scale Scores: Litr S502 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Reading	0.50	828.10	0.84
	Writing	0.50	2098.05	0.82
	Literacy			986.30
2-3	Reading	0.50	790.13	0.84
	Writing	0.50	2619.46	0.88
	Literacy			1175.12
4-5	Reading	0.50	1149.50	0.88
	Writing	0.50	2266.91	0.80
	Literacy			1307.76
6-8	Reading	0.50	1020.74	0.87
	Writing	0.50	1018.13	0.79
	Literacy			773.68
9-12	Reading	0.50	954.72	0.87
	Writing	0.50	1024.68	0.72
	Literacy			645.74

5.5.3 Comprehension

Table 5.5.3.1

Reliabilities of Composite Scale Scores: Cphn S502 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.30	3035.78	0.87
	Reading	0.70	1115.27	0.88
	Comprehension			1138.78
2-3	Listening	0.30	3523.26	0.86
	Reading	0.70	940.55	0.87
	Comprehension			1199.21
4-5	Listening	0.30	2400.06	0.82
	Reading	0.70	1232.06	0.90
	Comprehension			1271.60
6-8	Listening	0.30	2116.22	0.84
	Reading	0.70	1381.32	0.90
	Comprehension			1333.68
9-12	Listening	0.30	2731.05	0.87
	Reading	0.70	1339.30	0.91
	Comprehension			1468.22

Table 5.5.3.2

Reliabilities of Composite Scale Scores: Cphn S502 Online by Gender

Cluster	Component	Weight	Female		Male	
			Variance	Reliability	Variance	Reliability
1	Listening	0.30	2895.55	0.86	3123.85	0.87
	Reading	0.70	1157.01	0.89	1070.48	0.88
	Comprehension			1145.75	0.91	1119.38
2-3	Listening	0.30	3306.82	0.85	3688.97	0.86
	Reading	0.70	931.51	0.87	943.47	0.87
	Comprehension			1165.30	0.91	1220.56
4-5	Listening	0.30	2220.31	0.81	2527.81	0.82
	Reading	0.70	1153.31	0.89	1284.99	0.90
	Comprehension			1184.58	0.92	1334.56
6-8	Listening	0.30	2003.33	0.84	2192.58	0.85
	Reading	0.70	1335.66	0.90	1403.38	0.91
	Comprehension			1288.09	0.93	1359.96
9-12	Listening	0.30	2630.65	0.86	2791.66	0.87
	Reading	0.70	1293.46	0.90	1360.44	0.91
	Comprehension			1424.62	0.93	1489.93

Table 5.5.3.3

Reliabilities of Composite Scale Scores: Cphn S502 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.30	2903.93	0.87	3122.55	0.86
	Reading	0.70	858.77	0.85	1413.82	0.90
	Comprehension			899.21	0.89	1420.50
2-3	Listening	0.30	3443.16	0.86	3469.19	0.84
	Reading	0.70	851.72	0.86	1036.37	0.88
	Comprehension			1099.32	0.91	1289.38
4-5	Listening	0.30	2342.82	0.82	2256.92	0.80
	Reading	0.70	1169.90	0.89	1291.10	0.90
	Comprehension			1208.71	0.92	1295.44
6-8	Listening	0.30	2054.43	0.84	2087.35	0.83
	Reading	0.70	1298.25	0.90	1503.52	0.91
	Comprehension			1255.77	0.93	1418.93
9-12	Listening	0.30	2702.41	0.87	2550.36	0.85
	Reading	0.70	1268.08	0.90	1393.66	0.91
	Comprehension			1406.35	0.93	1466.00

Table 5.5.3.4

Reliabilities of Composite Scale Scores: Cphn S502 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.30	3203.66	0.89
	Reading	0.70	828.10	0.84
	Comprehension			908.57
2-3	Listening	0.30	3591.65	0.88
	Reading	0.70	790.13	0.84
	Comprehension			1023.66
4-5	Listening	0.30	2183.79	0.82
	Reading	0.70	1149.50	0.88
	Comprehension			1109.01
6-8	Listening	0.30	1691.47	0.81
	Reading	0.70	1020.74	0.87
	Comprehension			943.37
9-12	Listening	0.30	1888.93	0.82
	Reading	0.70	954.72	0.87
	Comprehension			953.60

5.5.4 Overall

Table 5.5.4.1

Reliabilities of Composite Scale Scores: Over S502 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.15	3035.78	0.87
	Reading	0.35	1115.27	0.88
	Writing	0.35	1716.10	0.78
	Speaking	0.15	2789.67	0.82
	Overall Composite			1072.34
2-3	Listening	0.15	3523.26	0.86
	Reading	0.35	940.55	0.87
	Writing	0.35	2166.15	0.84
	Speaking	0.15	2839.14	0.82
	Overall Composite			1287.27
4-5	Listening	0.15	2400.06	0.82
	Reading	0.35	1232.06	0.90
	Writing	0.35	2115.64	0.74
	Speaking	0.15	2416.71	0.83
	Overall Composite			1294.68
6-8	Listening	0.15	2116.22	0.84
	Reading	0.35	1381.32	0.90
	Writing	0.35	1369.06	0.77
	Speaking	0.15	2639.83	0.84
	Overall Composite			1147.22
9-12	Listening	0.15	2731.05	0.87
	Reading	0.35	1339.30	0.91
	Writing	0.35	1416.15	0.74
	Speaking	0.15	2978.28	0.83
	Overall Composite			1218.29

Table 5.5.4.2

Reliabilities of Composite Scale Scores: Over S502 Online by Gender

Cluster	Component	Weight	Female		Male	
			Variance	Reliability	Variance	Reliability
1	Listening	0.15	2895.55	0.86	3123.85	0.87
	Reading	0.35	1157.01	0.89	1070.48	0.88
	Writing	0.35	1544.24	0.77	1829.14	0.79
	Speaking	0.15	2783.12	0.82	2700.03	0.82
	Overall Composite		1034.11	0.92	1076.72	0.92
2-3	Listening	0.15	3306.82	0.85	3688.97	0.86
	Reading	0.35	931.51	0.87	943.47	0.87
	Writing	0.35	1989.64	0.82	2243.42	0.85
	Speaking	0.15	2779.17	0.82	2799.85	0.82
	Overall Composite		1221.70	0.93	1309.23	0.94
4-5	Listening	0.15	2220.31	0.81	2527.81	0.82
	Reading	0.35	1153.31	0.89	1284.99	0.90
	Writing	0.35	1846.90	0.71	2264.83	0.76
	Speaking	0.15	2363.33	0.82	2427.28	0.83
	Overall Composite		1178.77	0.91	1362.96	0.92
6-8	Listening	0.15	2003.33	0.84	2192.58	0.85
	Reading	0.35	1335.66	0.90	1403.38	0.91
	Writing	0.35	1341.14	0.75	1344.59	0.78
	Speaking	0.15	2670.55	0.84	2597.14	0.84
	Overall Composite		1123.74	0.93	1145.94	0.94
9-12	Listening	0.15	2630.65	0.86	2791.66	0.87
	Reading	0.35	1293.46	0.90	1360.44	0.91
	Writing	0.35	1393.59	0.73	1389.63	0.75
	Speaking	0.15	2940.86	0.83	2988.70	0.84
	Overall Composite		1200.27	0.93	1209.89	0.94

Table 5.5.4.3

Reliabilities of Composite Scale Scores: Over S502 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.15	2903.93	0.87	3122.55	0.86
	Reading	0.35	858.77	0.85	1413.82	0.90
	Writing	0.35	1605.82	0.79	1706.31	0.76
	Speaking	0.15	2730.26	0.82	2682.23	0.81
	Overall Composite		894.25	0.91	1217.51	0.93
2-3	Listening	0.15	3443.16	0.86	3469.19	0.84
	Reading	0.35	851.72	0.86	1036.37	0.88
	Writing	0.35	2288.70	0.85	1719.11	0.81
	Speaking	0.15	2872.60	0.82	2618.21	0.81
	Overall Composite		1265.59	0.94	1189.16	0.93
4-5	Listening	0.15	2342.82	0.82	2256.92	0.80
	Reading	0.35	1169.90	0.89	1291.10	0.90
	Writing	0.35	2127.81	0.74	1792.12	0.72
	Speaking	0.15	2396.52	0.83	2200.27	0.81
	Overall Composite		1261.38	0.92	1185.39	0.92
6-8	Listening	0.15	2054.43	0.84	2087.35	0.83
	Reading	0.35	1298.25	0.90	1503.52	0.91
	Writing	0.35	1354.40	0.78	1239.60	0.76
	Speaking	0.15	2636.49	0.84	2315.36	0.82
	Overall Composite		1100.85	0.94	1113.23	0.94
9-12	Listening	0.15	2702.41	0.87	2550.36	0.85
	Reading	0.35	1268.08	0.90	1393.66	0.91
	Writing	0.35	1405.12	0.74	1285.76	0.73
	Speaking	0.15	3001.12	0.84	2598.11	0.81
	Overall Composite		1186.34	0.93	1120.07	0.93

Table 5.5.4.4

Reliabilities of Composite Scale Scores: Over S502 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.15	3203.66	0.89
	Reading	0.35	828.10	0.84
	Writing	0.35	2098.05	0.82
	Speaking	0.15	3037.48	0.83
	Overall Composite			1010.63
2-3	Listening	0.15	3591.65	0.88
	Reading	0.35	790.13	0.84
	Writing	0.35	2619.46	0.88
	Speaking	0.15	3070.38	0.81
	Overall Composite			1277.46
4-5	Listening	0.15	2183.79	0.82
	Reading	0.35	1149.50	0.88
	Writing	0.35	2266.91	0.80
	Speaking	0.15	2242.96	0.81
	Overall Composite			1160.90
6-8	Listening	0.15	1691.47	0.81
	Reading	0.35	1020.74	0.87
	Writing	0.35	1018.13	0.79
	Speaking	0.15	2263.24	0.83
	Overall Composite			766.20
9-12	Listening	0.15	1888.93	0.82
	Reading	0.35	954.72	0.87
	Writing	0.35	1024.68	0.72
	Speaking	0.15	2530.12	0.83
	Overall Composite			711.07

5.6 Conditional Standard Errors of Measurement 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 subtest 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 use a three-step process to derive the CSEM for each ACCESS composite scale score. We calculate 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 use 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, are independent of each other. Therefore, we can estimate the CSEM for a student's composite scale score x , SEM_x , using the equation derived by Price et al. (2006):

$$SEM_x = \sqrt{W_1^2 SEM_1^2 + W_2^2 SEM_2^2 + W_3^2 SEM_3^2 + \dots + W_k^2 SEM_k^2}$$

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 $i=1, \dots, k$.

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 were not observed in the data.

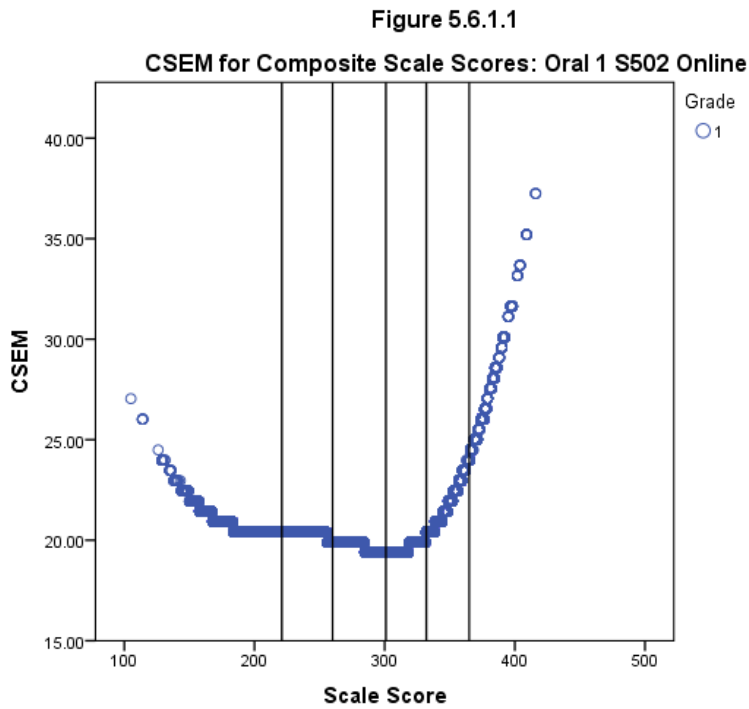
The figures in this section show graphically the CSEMs for various composite scale scores by grade level. The students' composite scale scores appear on the horizontal axis, and the corresponding CSEMs appear on the vertical axis. Each point in a figure represents a student in the dataset, showing the relationship between the CSEM and that student's composite scale score. We do 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.

Five vertical lines in the figure indicate the five ACCESS cut points for the highest grade in the grade-level cluster for the test form, dividing the figure into six sections for each of the WIDA proficiency levels (1–6) for the composites.

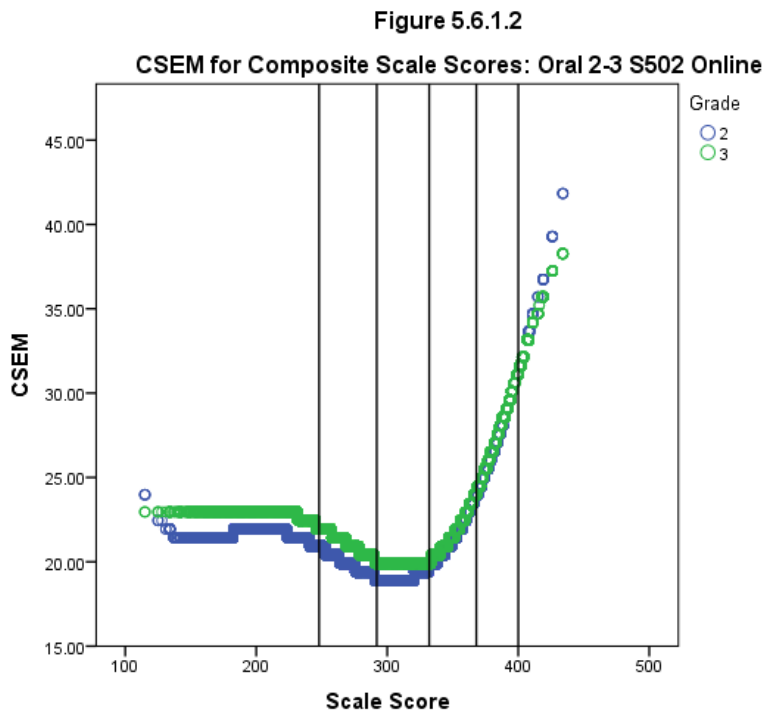
Low CSEM values indicate less measurement error (i.e., greater accuracy in measurement). In general, these figures show that the CSEMs are lower and fairly constant in the middle of the composite scale score range and higher and more variable for extreme low and high composite scale scores. This is to be expected, as ACCESS test items and scores were scaled using the IRT method, which is known to produce higher CSEMs at the lower and the higher end of the scale score range. In addition, because students exit the EL program when they demonstrate that they are English language proficient, the numbers of students at the extreme high composite scale score range are typically small as compared to those at the middle composite scale score range. Therefore, the measurement errors associated with the scale scores at the extreme high composite scale score range tend to be higher since there are fewer students available in estimating the scores and the measurement errors for these scores.

5.6.1 Oral

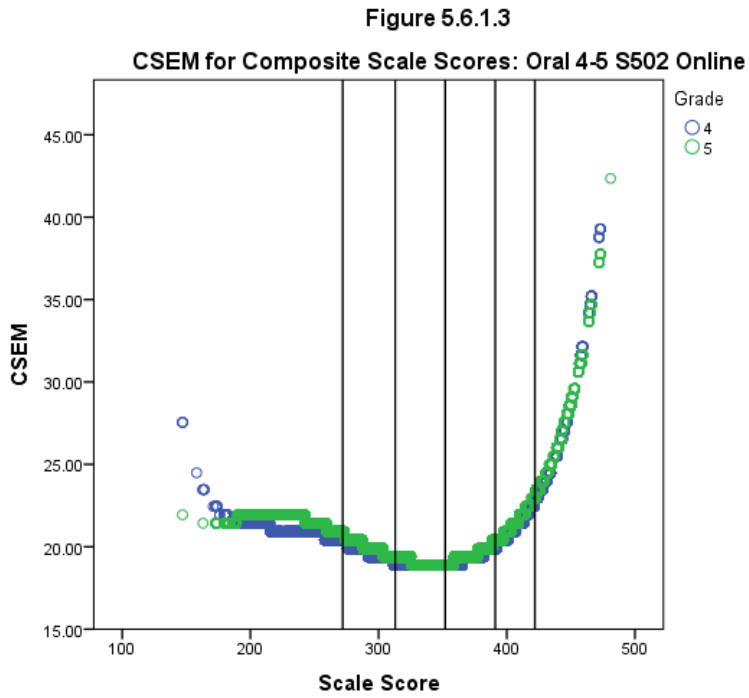
5.6.1.1 Grade 1



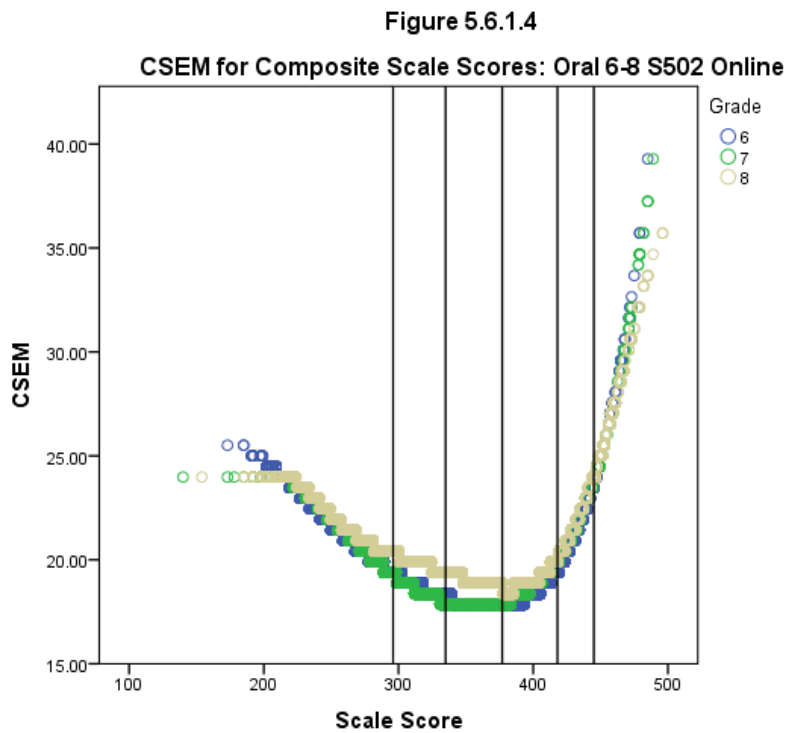
5.6.1.2 Grades 2-3



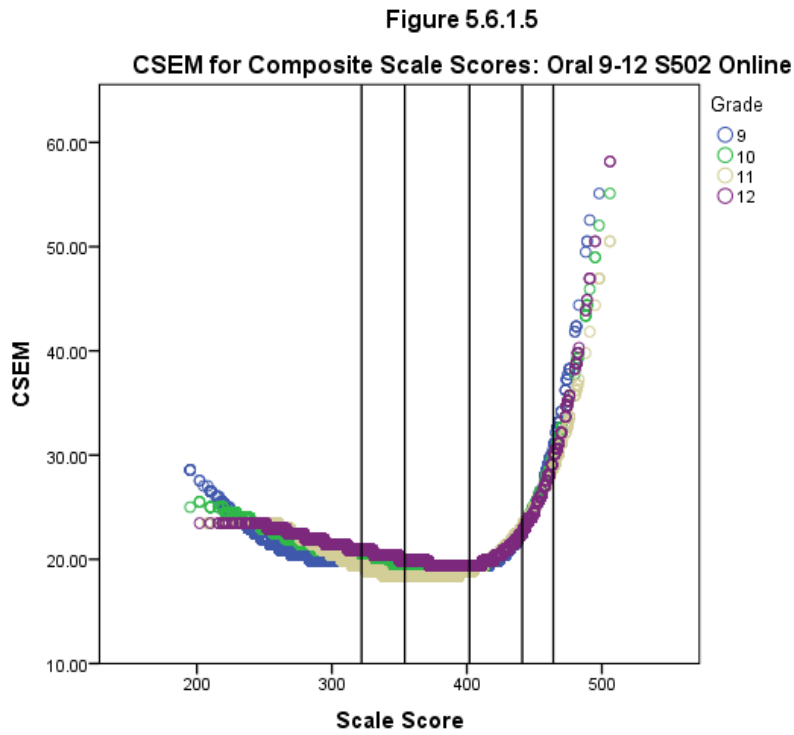
5.6.1.3 Grades 4–5



5.6.1.4 Grades 6–8



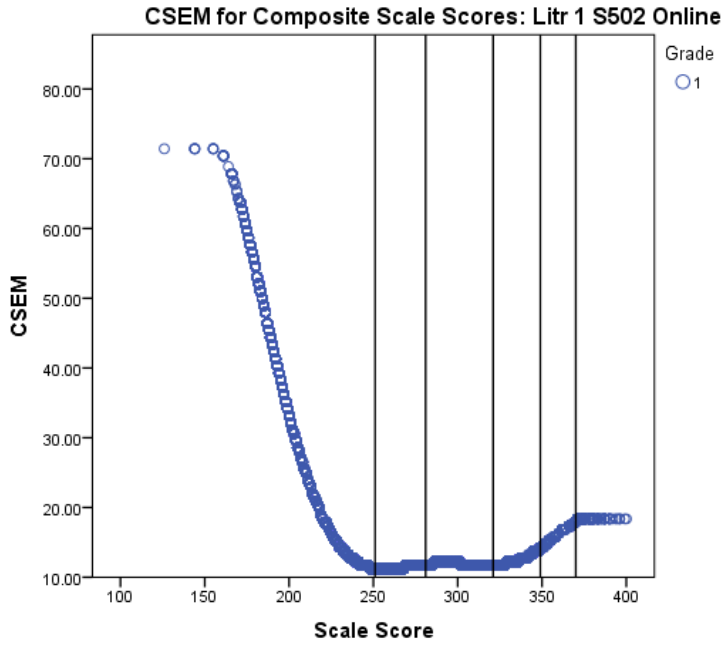
5.6.1.5 Grades 9-12



5.6.2 Literacy

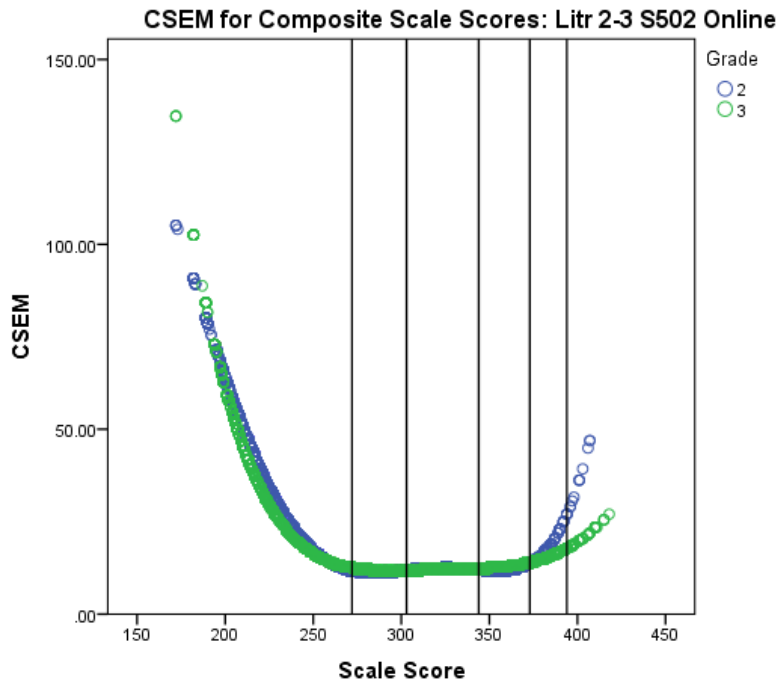
5.6.2.1 Grade 1

Figure 5.6.2.1

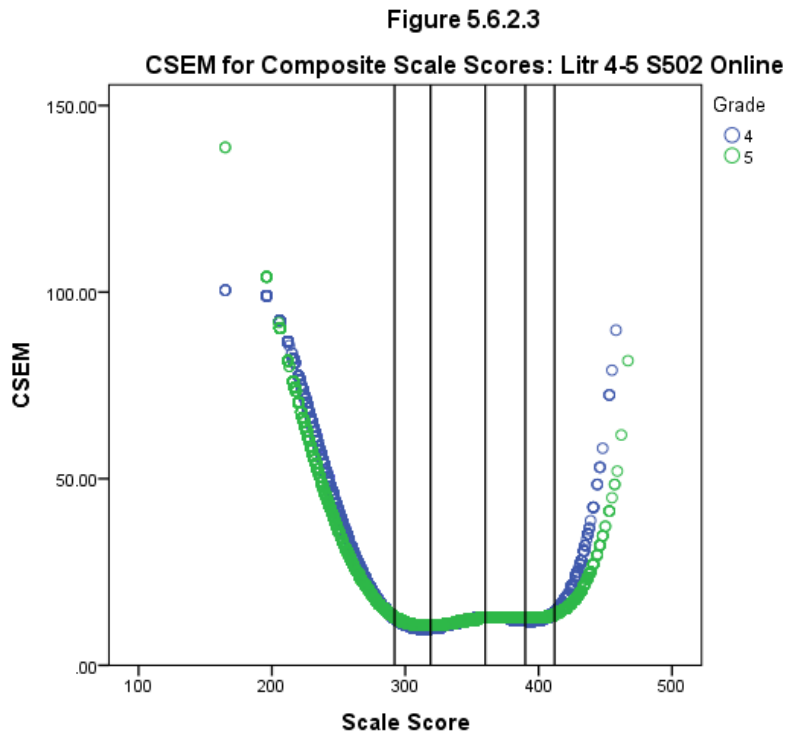


5.6.2.2 Grades 2-3

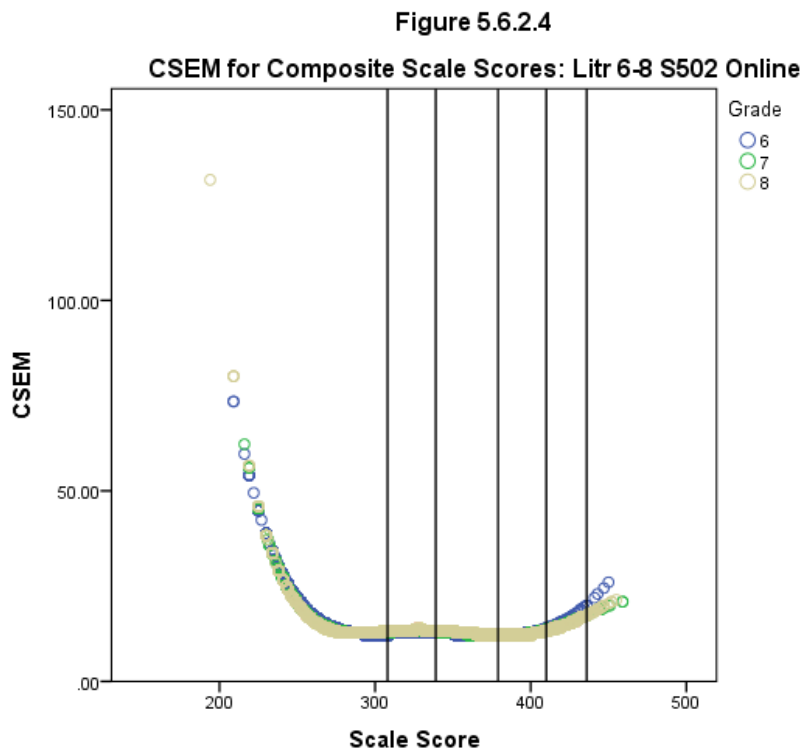
Figure 5.6.2.2



5.6.2.3 Grades 4–5

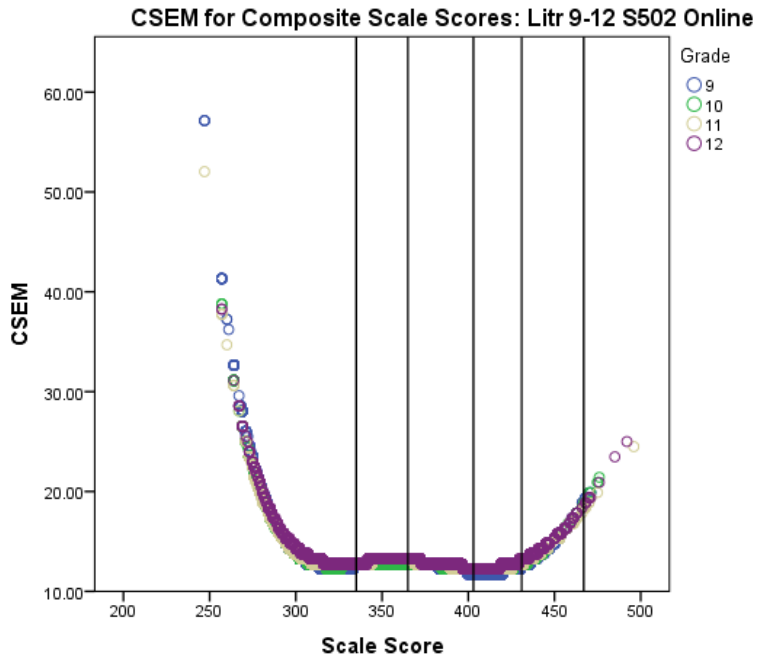


5.6.2.4 Grades 6–8



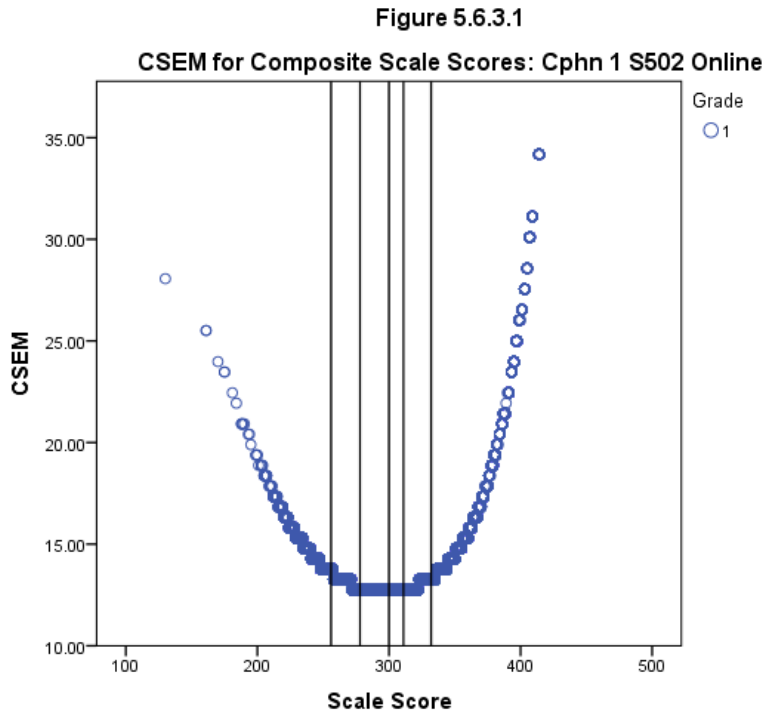
5.6.2.5 Grades 9-12

Figure 5.6.2.5

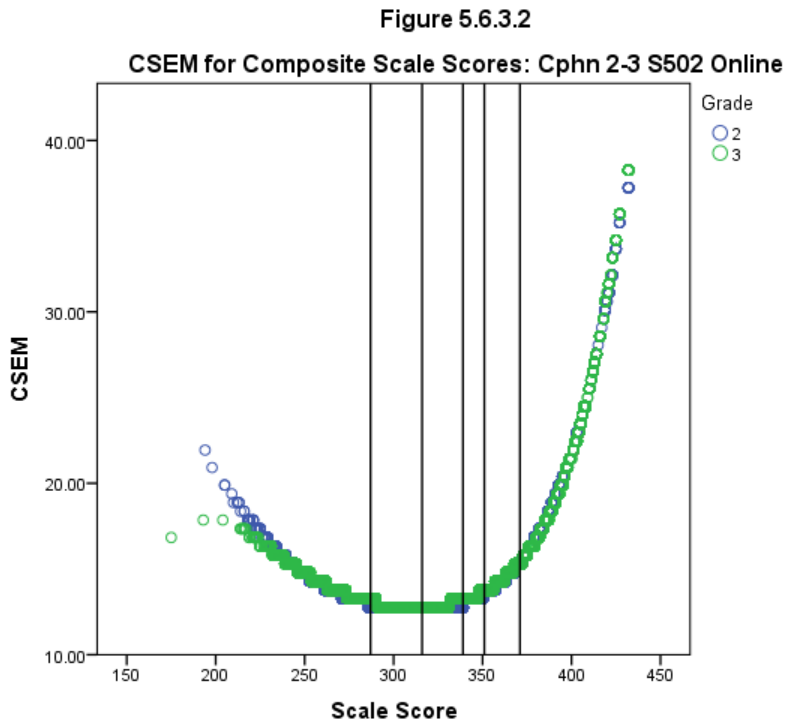


5.6.3 Comprehension

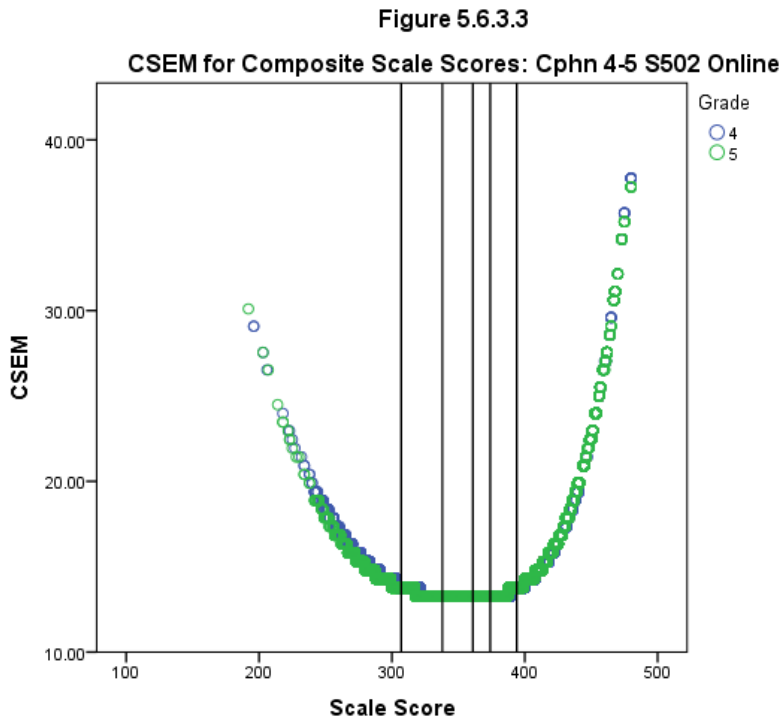
5.6.3.1 Grade 1



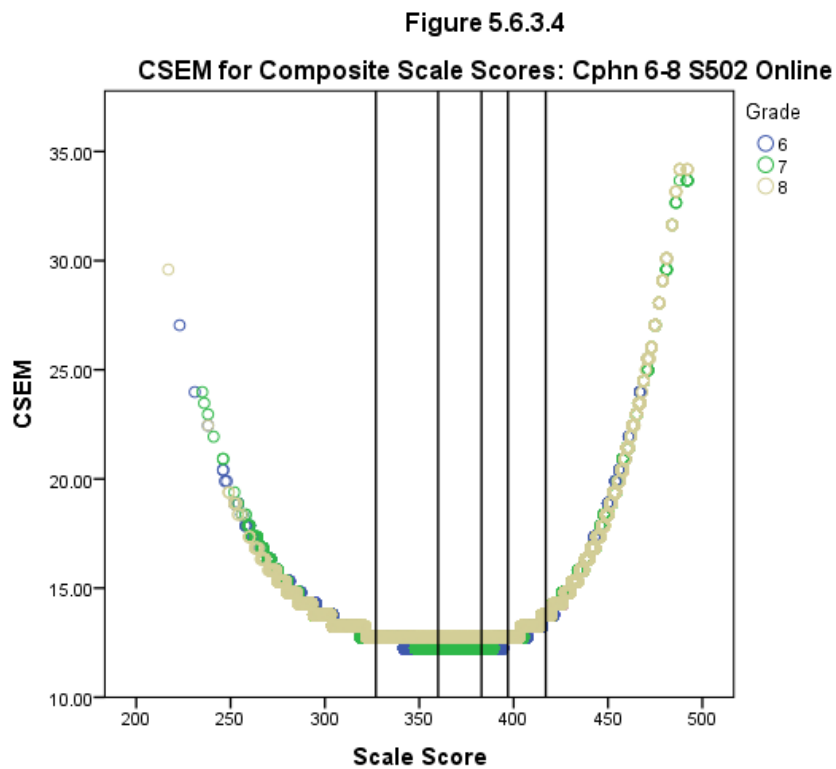
5.6.3.2 Grades 2-3



5.6.3.3 Grades 4–5

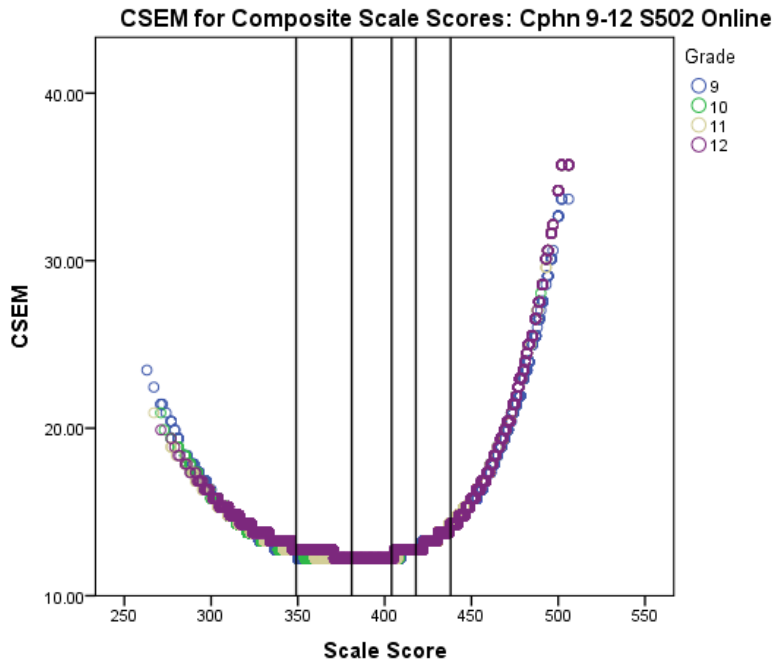


5.6.3.4 Grades 6–8



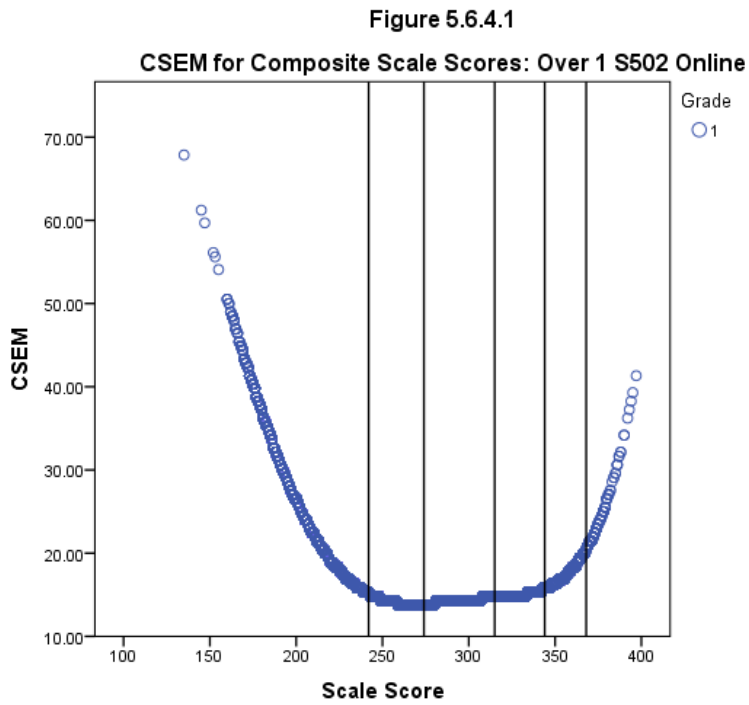
5.6.3.5 Grades 9-12

Figure 5.6.3.5

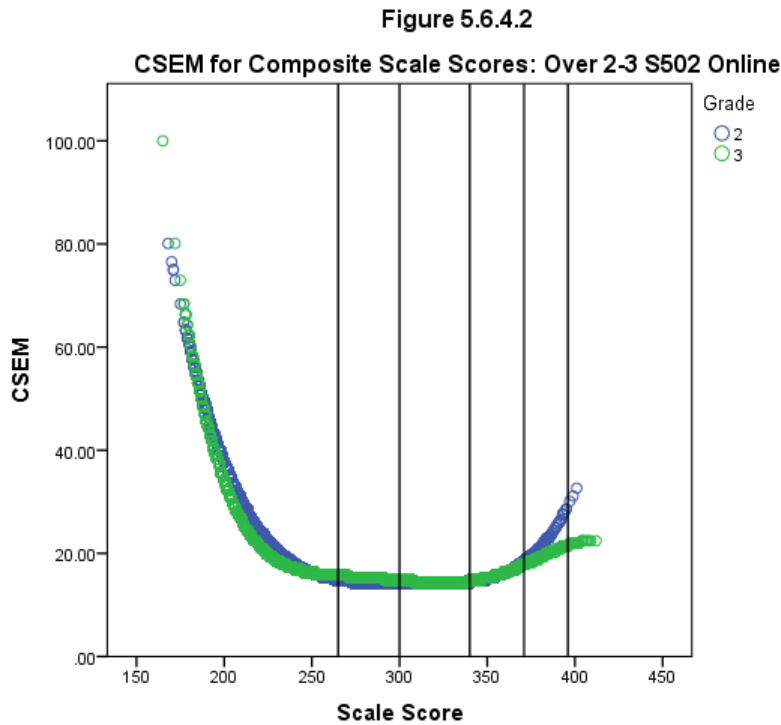


5.6.4 Overall

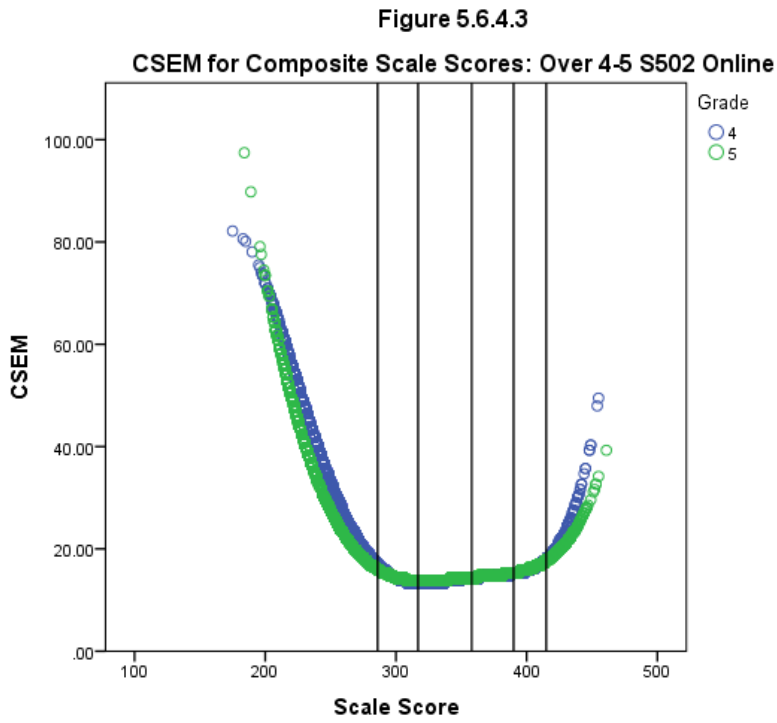
5.6.4.1 Grade 1



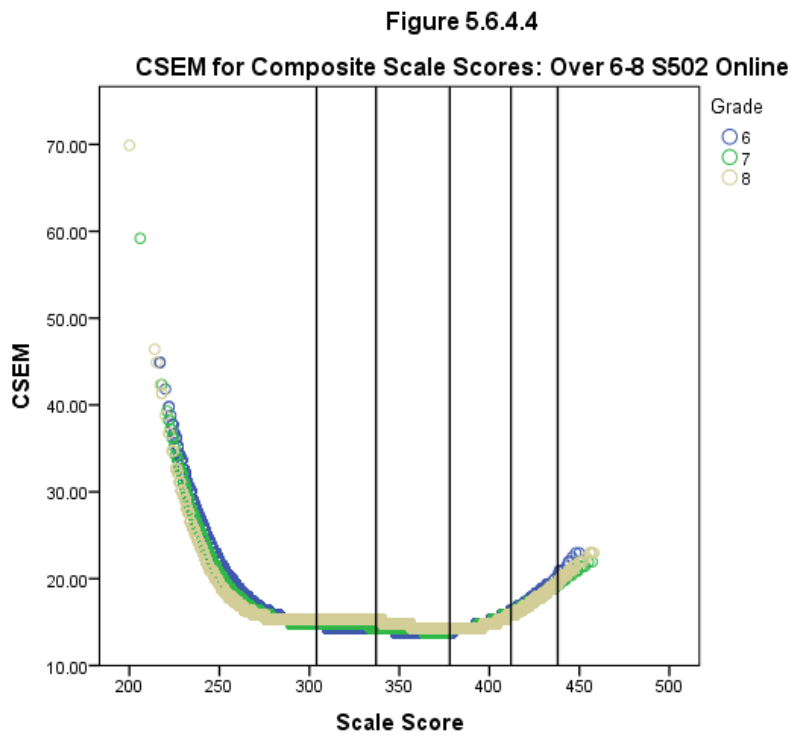
5.6.4.2 Grades 2-3



5.6.4.3 Grades 4–5

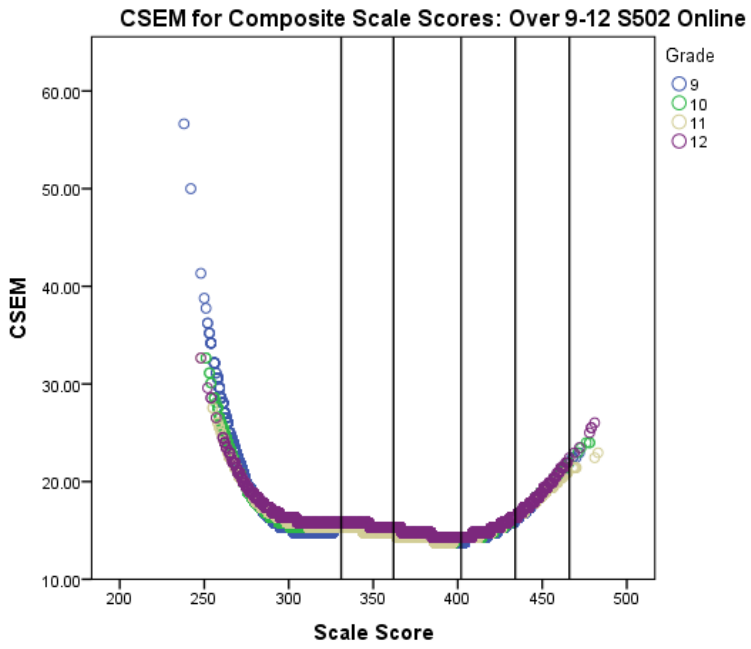


5.6.4.4 Grades 6–8



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 test domain, 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 scale scores at the cut points for each grade. The third reports the marginal classification consistency indices based on the 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 the range of these indices, by each composite, highlighting the grade with the lowest classification accuracy and consistency indices for each 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 0.674 to 0.743, and the overall classification consistency indices ranged from 0.564 to 0.649 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 0.690 to 0.797, and the overall classification consistency indices ranged from 0.584 to 0.715 across grades. The lowest overall classification accuracy and consistency indices were found for students in Grade 5.

For the Comprehension composite, as shown in Table 5.7.3.1, the overall classification accuracy indices ranged from 0.637 to 0.707, and the overall classification consistency indices ranged from 0.531 to 0.610 across grades. The lowest overall classification accuracy and consistency indices were found for students in Grade 1.

For the Overall composite, as shown in Table 5.7.4.1, the overall classification accuracy indices ranged from 0.750 to 0.838, and the overall classification consistency indices ranged from 0.663 to 0.773 across grades. The lowest overall classification accuracy and consistency indices were found for students in Grade 5.

The results suggest that the lowest overall classification accuracy and consistency indices for three out of the four composites (Oral, Literacy, and Overall) were found for students in Grade 5, while the lowest overall classification accuracy and consistency indices for the Comprehension composite were found for students in Grade 1.

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 the range of the marginal classification accuracy and consistency indices for the composite scale scores across grades and highlight the grade with the lowest marginal classification accuracy and the lowest consistency indices, by composite.

For the Oral composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.870 to 0.995 (Table 5.7.1.2), and the marginal classification consistency indices ranged from 0.818 to 0.995 (Table 5.7.1.3). The lowest marginal classification accuracy and consistency indices were found for students in Grade 5, at the PL 4/PL 5 cut point. 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/PL 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 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 0.867 to 0.999 (Table 5.7.2.2), and the marginal classification consistency indices ranged from 0.814 to 0.998 (Table 5.7.2.3). The lowest marginal classification accuracy and consistency indices were found for students in Grade 4, at the PL 3/PL 4 cut point. Note that Grade 4 also had the lowest overall classification accuracy and consistency indices for the Literacy composite followed by Grade 5. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut point appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal and overall accuracy and consistency indices for the Grades 4 and 5 Literacy composite are still in the 0.80 to 0.90 range.

For the Comprehension composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.899 to 0.985 (Table 5.7.3.2), and the marginal classification consistency indices ranged from 0.859 to 0.978 (Table 5.7.3.3). The lowest marginal classification accuracy and consistency indices were found for students in Grade 1, at the PL 2/PL 3 cut point. 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/PL 3 cut point appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal and overall accuracy and consistency indices for Grade 1 Comprehension are still in the high 0.80 range and the mid 0.90 range.

For the Overall composite, the marginal classification accuracy indices based on the scale scores at the cut points ranged from 0.905 to 0.999 (Table 5.7.4.2), and the marginal classification consistency indices ranged from 0.867 to 0.999 (Table 5.7.4.3). The lowest marginal classification accuracy and consistency indices were found for students in Grade 5, at the PL 3/PL 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 3/PL 4 cut point appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal and overall accuracy and consistency indices for the Grade 5 Overall composite are still in the 0.80 to 0.90 range.

The overall and marginal classification accuracy and consistency indices provided similar findings. That is, as overall classification accuracy and consistency indices were lower, the marginal classification accuracy and consistency indices tended to be lower. Especially, the lowest overall and marginal classification accuracy and consistency indices for two of the four composites (Oral and Overall) were found for students in Grade 5, while 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. Marginal classification accuracy and consistency are expected to vary for different ability levels due to variation in measurement accuracy. The further away a student's scale score is from the cut point, the smaller the classification errors would be, or the more accurate the classification decision would be. With a large number of proficiency levels, there are more student scale scores that are near the cut points than there would be 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). Since ACCESS has six proficiency levels and some proficiency levels such as PL 3 and PL 4 occupy relatively narrow ranges on the ability scale compared with other proficiency levels, the marginal classification accuracy and consistency indices based on scale scores for cut points in the middle ranges are lower than for other cut points.

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 use 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 0.814 for all composite scale scores. Additionally, the marginal classification accuracy and consistency indices were at or above 0.867 for the Overall composite scale scores, where test users make most of the accountability decisions.

5.7.1 Oral

Table 5.7.1.1

Overall Accuracy and Consistency of Classification Indices: Oral S502 Online

Grade	Accuracy	Consistency
1	0.685	0.576
2	0.704	0.597
3	0.686	0.578
4	0.678	0.570
5	0.674	0.564
6	0.742	0.645
7	0.726	0.623
8	0.706	0.602
9	0.741	0.642
10	0.734	0.635
11	0.735	0.637
12	0.743	0.649

Table 5.7.1.2

Classification Accuracy Indices at Cut Score Level: Oral S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.959	0.921	0.897	0.925	0.981
2	0.955	0.910	0.902	0.943	0.992
3	0.962	0.923	0.887	0.917	0.993
4	0.988	0.965	0.906	0.877	0.942
5	0.982	0.955	0.898	0.870	0.966
6	0.981	0.943	0.891	0.934	0.992
7	0.973	0.932	0.891	0.941	0.988
8	0.967	0.927	0.890	0.931	0.987
9	0.954	0.917	0.904	0.968	0.995
10	0.944	0.913	0.909	0.969	0.995
11	0.948	0.913	0.907	0.967	0.995
12	0.944	0.911	0.912	0.975	N/A

Table 5.7.1.3

Classification Consistency Indices at Cut Score Level: Oral S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.942	0.888	0.856	0.893	0.975
2	0.936	0.873	0.863	0.920	0.991
3	0.946	0.891	0.841	0.887	0.992
4	0.983	0.948	0.868	0.826	0.928
5	0.975	0.934	0.858	0.818	0.957
6	0.973	0.918	0.847	0.909	0.992
7	0.961	0.903	0.847	0.915	0.987
8	0.953	0.896	0.846	0.906	0.985
9	0.935	0.882	0.865	0.956	0.995
10	0.921	0.877	0.872	0.957	0.995
11	0.925	0.877	0.868	0.959	0.995
12	0.920	0.875	0.875	0.968	N/A

5.7.2 Literacy

Table 5.7.2.1

Overall Accuracy and Consistency of Classification Indices: Litr S502 Online

Grade	Accuracy	Consistency
1	0.762	0.670
2	0.769	0.679
3	0.757	0.667
4	0.700	0.594
5	0.690	0.584
6	0.797	0.715
7	0.777	0.689
8	0.761	0.668
9	0.751	0.654
10	0.747	0.647
11	0.752	0.653
12	0.756	0.660

Table 5.7.2.2

Classification Accuracy Indices at Cut Score Level: Litr S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.891	0.906	0.974	0.993	0.999
2	0.958	0.908	0.913	0.990	N/A
3	0.967	0.924	0.886	0.979	N/A
4	0.969	0.938	0.867	0.925	0.986
5	0.968	0.935	0.870	0.915	0.984
6	0.952	0.904	0.947	0.994	N/A
7	0.951	0.906	0.933	0.987	N/A
8	0.943	0.903	0.929	0.986	N/A
9	0.963	0.913	0.906	0.969	N/A
10	0.954	0.903	0.916	0.974	N/A
11	0.956	0.901	0.918	0.977	N/A
12	0.950	0.896	0.927	0.983	N/A

Table 5.7.2.3

Classification Consistency Indices at Cut Score Level: Litr S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.846	0.865	0.965	0.991	0.998
2	0.939	0.871	0.877	0.989	N/A
3	0.953	0.893	0.840	0.975	N/A
4	0.956	0.909	0.814	0.901	0.985
5	0.954	0.906	0.819	0.893	0.982
6	0.931	0.864	0.923	0.994	N/A
7	0.931	0.867	0.903	0.985	N/A
8	0.919	0.864	0.900	0.981	N/A
9	0.947	0.876	0.867	0.959	N/A
10	0.935	0.863	0.880	0.964	N/A
11	0.937	0.861	0.884	0.967	N/A
12	0.928	0.855	0.898	0.975	N/A

5.7.3 Comprehension

Table 5.7.3.1

Overall Accuracy and Consistency of Classification Indices: Cphn S502 Online

Grade	Accuracy	Consistency
1	0.637	0.531
2	0.674	0.568
3	0.660	0.557
4	0.701	0.607
5	0.680	0.583
6	0.699	0.598
7	0.681	0.580
8	0.677	0.577
9	0.707	0.610
10	0.702	0.606
11	0.702	0.606
12	0.703	0.605

Table 5.7.3.2

Classification Accuracy Indices at Cut Score Level: Cphn S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.928	0.899	0.908	0.927	0.958
2	0.965	0.915	0.908	0.922	0.954
3	0.952	0.914	0.910	0.920	0.946
4	0.985	0.955	0.928	0.910	0.913
5	0.976	0.949	0.919	0.904	0.917
6	0.966	0.924	0.908	0.928	0.967
7	0.960	0.923	0.907	0.926	0.956
8	0.952	0.920	0.912	0.928	0.953
9	0.968	0.933	0.922	0.927	0.950
10	0.959	0.931	0.925	0.928	0.951
11	0.960	0.931	0.924	0.928	0.951
12	0.957	0.929	0.926	0.929	0.955

Table 5.7.3.3

Classification Consistency Indices at Cut Score Level: Cphn S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.902	0.859	0.872	0.896	0.940
2	0.950	0.881	0.871	0.890	0.935
3	0.932	0.879	0.874	0.888	0.924
4	0.978	0.936	0.897	0.874	0.877
5	0.967	0.927	0.885	0.868	0.883
6	0.953	0.893	0.873	0.898	0.953
7	0.944	0.891	0.872	0.895	0.937
8	0.932	0.887	0.879	0.897	0.933
9	0.955	0.905	0.890	0.898	0.930
10	0.943	0.902	0.894	0.900	0.930
11	0.944	0.902	0.893	0.899	0.930
12	0.940	0.900	0.895	0.901	0.935

5.7.4 Overall

Table 5.7.4.1

Overall Accuracy and Consistency of Classification Indices: Over S502 Online

Grade	Accuracy	Consistency
1	0.805	0.725
2	0.815	0.743
3	0.803	0.729
4	0.754	0.666
5	0.75	0.663
6	0.838	0.773
7	0.819	0.749
8	0.805	0.732
9	0.809	0.737
10	0.803	0.726
11	0.810	0.734
12	0.815	0.740

Table 5.7.4.2

Classification Accuracy Indices at Cut Score Level: Over S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.940	0.909	0.966	0.991	0.999
2	0.969	0.929	0.934	0.984	N/A
3	0.975	0.941	0.910	0.977	N/A
4	0.983	0.960	0.908	0.913	0.986
5	0.981	0.955	0.905	0.912	0.989
6	0.976	0.936	0.933	0.991	N/A
7	0.972	0.935	0.930	0.983	N/A
8	0.966	0.931	0.930	0.979	N/A
9	0.971	0.936	0.927	0.975	N/A
10	0.963	0.931	0.932	0.977	N/A
11	0.964	0.929	0.936	0.982	N/A
12	0.960	0.925	0.943	0.987	N/A

Table 5.7.4.3

Classification Consistency Indices at Cut Score Level: Over S502 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.913	0.872	0.952	0.989	0.999
2	0.955	0.900	0.906	0.982	N/A
3	0.964	0.916	0.873	0.975	N/A
4	0.976	0.942	0.871	0.883	0.986
5	0.973	0.936	0.867	0.889	0.989
6	0.966	0.910	0.905	0.992	N/A
7	0.960	0.907	0.900	0.982	N/A
8	0.952	0.903	0.900	0.977	N/A
9	0.959	0.909	0.897	0.971	N/A
10	0.948	0.902	0.903	0.972	N/A
11	0.949	0.900	0.909	0.977	N/A
12	0.944	0.894	0.919	0.984	N/A

6 Quality Control

6.1 Content Development Quality Control

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 resume 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 and 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 and 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 and Sensitivity review focuses on selecting educators with culturally and linguistically diverse backgrounds who have experience interacting with ELs from a range of cultural, regional, religious, linguistic, ethnic, and socioeconomic backgrounds.

At the beginning of both Bias and 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 and 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 and Sensitivity review of Grade 1 and Grades 2–3 is often combined). Although Bias and 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 and 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 tryout 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://grow.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 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 upcoming 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(es) 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 quality control, please refer to Section 4.2.

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.³ This team examines data from three districts: a primary district, for

³ Due to the COVID-19 pandemic, the 2021 Score Reporting quality control was conducted online, with only WIDA and DRC employees participating.

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 Report to check the District Frequency Report. Finally, panelists check the State Frequency Report 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 students plagiarizing responses on the Grades 9–12 Speaking and/or Writing tests. 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), either via translation software or screen reading functionality.

It is likely that in addition to student malpractice, there may be issues with how Test Administrators were monitoring these test sessions. Since students are not able to access the internet from the actual testing device when INSIGHT is in use, an unapproved device such as a cell phone must have been used to access the websites. WIDA is concerned about both the validity

of individual students' scores as a result of the malpractice and the security of ACCESS test items. In all identified cases, state departments of education staff were notified shortly after discovery.

Consistent with the scoring process established last year, all responses containing plagiarized content will receive a nonscorable code of "Invalid Indecipherable," and a plagiarized flag in DRC's scoring platform is applied. This means that the student will receive a score of zero for tasks with plagiarized responses. Tasks that students respond to without plagiarizing will be scored as normal. States may then decide to use the score or invalidate the score in their own systems.

Below is the summary of the cases of plagiarism in Speaking and Writing domains by state.

State	Student Counts		
	Speaking	Writing	Speaking Responses Using an Artificial Voice
AL	1		
CO	2		
DE		1	
FL		8	
GA	3	1	1
ID		2	
IL	19	2	1
IN	2		1
KY	3		
MA	6	1	
MD	1		
ME	2		
MI	10		
MN	3	1	
MO	3		
MT	1		
NC	6	1	
ND	2		
NJ	5	1	
NM	1		1
NV	3		
OK	16	1	6
PA	10		
RI	1		
SC		2	
TN	4		2
UT	5		
VA	7		1
VT	1		
WI	9		1
Totals			
30 States	126 Students	21 Students	14 Students

Caveon Data Forensic Analysis Results

WIDA hired Caveon to perform data forensic analysis during the 2020–2021 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 2020 through August 2021 Online multistage adaptive test administrations, Listening and Reading domains
- December 2020 through August 2021 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 may 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 preknowledge of test content. By applying aberrance measures and comparing the performance between aberrant and nonaberrant 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 IRT 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

- Allen, N. L., Carlson, J. E., & Zalanak, C. A. (1999). *The NAEP 1996 technical report*. Washington, DC: 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*. Washington, DC: American Psychological Association.
- American Institutes of Research. (2018). *ELPA21 technical report, part I – summative assessment*. Washington, DC: 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]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.
- Center for Applied Linguistics. (2016). *ACCESS for ELLs® Series 400 Listening and Reading scale maintenance: Technical brief*. Washington, DC: Author.
- Center for Applied Linguistics. (2017). *ACCESS for ELLs® 2.0 Speaking and Writing score scale reconstruction: Technical brief*. Washington, DC: Author.
- Center for Applied Linguistics. (2019). *Maintaining the ACCESS for ELLs Online Writing Scale: Preparations for the Series 501 redesign: Technical brief*. Washington, DC: Author.
- Cook, H. G., & MacGregor, D. (2017). *The ACCESS for ELLs 2.0 2016 Standard setting study* [Technical Report]. Madison, WI: 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, Iowa City, IA. doi: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). New York: 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*. Madison, WI: WIDA Consortium.
- Jones, P. E., & Smith, R. W. (2006). *Item parameter drift in certification exams and its impact on pass-fail decision making*. Presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, CA.
- Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores*. 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]. Washington, DC: 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]. Washington, DC: Center for Applied Linguistics.
- Kim, A., Chapman, & M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy Required for interpreting score reports: A focus on educators of K-12 English learners, *Language Testing, Vol. 37*(1) 54-75.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.
- 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. Retrieved from <http://www.rasch.org/rmt/rmt132i.htm>
- Linacre, J. M. (2002a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878. Retrieved from <http://www.rasch.org/rmt/rmt162f.htm>.
- Linacre, J. M. (2002b). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85–106.
- Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. N. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press.
- Linacre, J. M. (2006). Winsteps Rasch analysis (Version 3.60.1) [Computer software]. Retrieved from <http://www.winsteps.com>
- Linacre, J. M. (2020). *Reliability and separation of measures*. Retrieved from <https://www.winsteps.com/winman/reliability.htm>
- Linacre, J. M. (n.d.). *Displacement measures*. Retrieved from <http://www.winsteps.com/winman/displacement.htm>
- Livingston, S. A. (2018). *Reliability—basic concepts* [ETS Research Memorandum No. RM-18-01]. Princeton, NJ: ETS.
- 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*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- MacGregor, D., Yen, S., & Yu, X. (2021). Using multistage testing to enhance measurement of an English language proficiency test. *Language Assessment Quarterly*. doi: 10.1080/15434303.2021.1988953
- 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]. Retrieved from <http://itemanalysis.com/jmetrik-download/>
- Min, S., Bishop, K., & Cook, H. G. (2021). Reading is a multidimensional construct at child-L2-English-literacy onset, but comprises fewer dimensions over time: Evidence from multidimensional IRT analysis. *Language Testing*. doi: 10.1177/02655322211045296

- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 4, 351–363.
- National Center on Educational Outcomes. (2021). *Universal design of assessments*. Retrieved from 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. September, 2020.
- 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). Mahwah, NJ: 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). Hillsdale, NJ: 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*. Retrieved from https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf?utm_content=&utm_medium=email&utm_name=&utm_source=govdelivery&utm_term=
- WIDA Consortium. (2007). *English Language Proficiency Standards and resource guide, 2007 edition, Pre-Kindergarten through Grade 12*. Madison, WI: Board of Regents of the University of Wisconsin System.
- WIDA Consortium. (2012). *2012 amplification of the English Language Development Standards Kindergarten–Grade 12*. Madison, WI: 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*. Madison, WI: Board of Regents of the University of Wisconsin System
- WIDA Consortium. (2021a). *ACCESS for ELLs Test Administrator manual*. Madison, WI: Board of Regents of the University of Wisconsin System.
- WIDA Consortium. (2021b). *ACCESS for ELLs district and school test coordinator manual*. Madison, WI: Board of Regents of the University of Wisconsin System.
- WIDA Consortium. (2021c). *Test policy handbook*. Madison, WI: Board of Regents of the University of Wisconsin System.
- WIDA Consortium. (2021d). *Individual student report 2021*. Retrieved from <https://wida.wisc.edu/sites/default/files/resource/ACCESS-Sample-Individual-Score-Report-English.pdf>
- 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., & Douglas, G. A. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97–116.
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: 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]. Los Angeles, CA: 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). Hillsdale, NJ: 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). Hoboken, NJ: 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]. Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1993.tb01522.x

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.
Yage (Leah) Guo, Ph.D.
Michele Kawood, M.S.Ed.
Justin Kelly, Ph.D.
Dorry M. Kenyon, Ph.D.
Jung-Jung Lee, M.Sc.
Isabella De Leon, B.S.
Erin Shaw-Meadow, M.Sc.
Samantha Musser, M.A.
Rachel Myers, M.S.
Yoon Ah Song, Ph.D.
Alice Tsai, M.S.
Frank Wucinski, M.A.
Shu Jing Yen, Ph.D.
Xin Yu, M.A.

From WIDA:

Anna Rhoad-Drogalis, MS.
Kyoungwon Bishop, Ph.D.
Sakine Göçer Sahin, Ph.D.