World-Class Instructional Design and Assessment

Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 501, 2019–2020 Administration

Annual Technical Report No. 16A

Prepared by:

Center for Applied Linguistics

Language Assessment Division
Psychometrics and Quantitative Research Team

May 2021

The WIDA ACCESS for ELLs Technical Advisory Committee

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

- Jamal Abedi, Ph.D., Professor, Graduate School of Education, University of California at Davis and a research partner at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
- Lyle Bachman, Ph.D., Professor Emeritus, Applied Linguistics, University of California, Los Angeles
- 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 16th annual technical report on the ACCESS for ELLs English Language Proficiency Test and the fifth report on the ACCESS for ELLs assessment given in Online format.

This technical report is produced as a service to members and potential members of the WIDA Consortium and to support states' submissions for U.S. Department of Education English language proficiency assessment peer review. The technical information herein is intended for use by those who have technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). WIDA also produces an annual *Year in Review Report*, intended for a general audience, for readers who are interested in a nontechnical overview of the 2019–2020 ACCESS assessment.

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

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

ACCESS for ELLs Online Series 501 was administered in school year 2019–2020 in 34 states, the Bureau of Indian Education, the District of Columbia, the Commonwealth of the Northern Mariana Islands, and the U.S. Virgin Islands, for a total of 38 state entities (henceforth "states").

The Series 501 Online data set included the results of 1,571,889 students. The largest grade was Grade 2 with 194,261students, while the smallest was Grade 12 with 62,369 students. Of the participating WIDA states, the largest was Illinois with 194,452 students, while the smallest was the U.S. Virgin Islands with 158 students.

During the 2019–2020 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 most students who likely would participate in ACCESS for ELLs had completed their test sessions at the time that schools closed. Further detail on the impact of COVID-19 is contained in the ACCESS 2019–2020 *Year in Review Report*.

ACCESS for ELLs Series 501 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 16A) provides technical information pertaining to ACCESS for ELLs Series 501 Online. A second report (WIDA ACCESS Technical Report 16B) provides technical information for the ACCESS for ELLs Series 501 Paper assessment, including the Kindergarten assessment.

Part 1:
Purpose, Design, Implementation

Contents

1. Purpose and Design of ACCESS

1.1. Purpose Statement

The purpose of ACCESS for ELLs is to assess the developing English language proficiency of English language learners (ELLs) in Grades K–12 in the United States as defined by the multistate 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 English language learners.

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 the cases of some test items and tasks, the Standards are combined as follows:

- Integrated Social and Instructional Language (SIL), Language of Language Arts (LoLA), and Language of Social Studies (LoSS): IT
- Language of Math (LoMA) and Language of Science (LoSC): MS
- Language of Language Arts (LoLA) and Language of Social Studies (LoSS): LS

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.

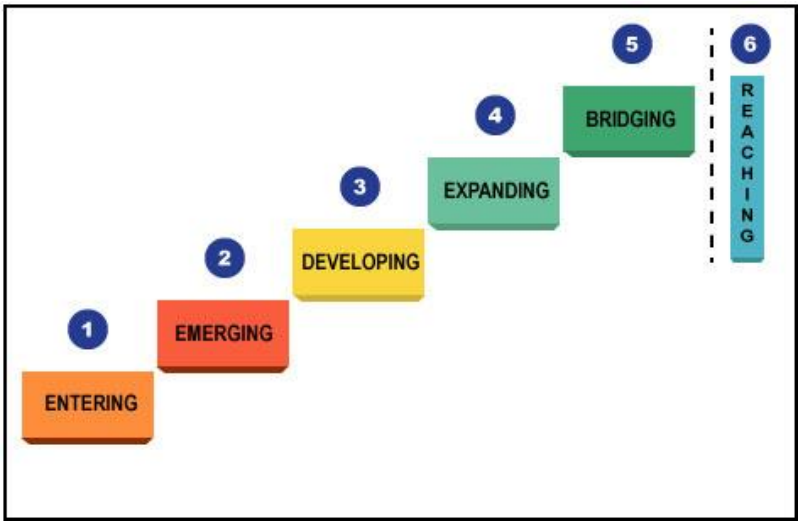


Figure 1. The language proficiency levels of the WIDA ELD Standards.

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.

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. Tests must be at the appropriate difficulty level for each individual test taker in order to be valid and reliable. While the grade-level cluster structure is a design feature intended to ensure that the language expectations are developmentally appropriate for children at 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 test takers may be boring and lead to inattentiveness on the part of students; items that are far too difficult for test takers 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. 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.

2. Test Development

2.1. Test Design

This section describes how ACCESS Online is assembled to ensure that the evidence collected is (a) sufficient to make the intended decisions, and (b) appropriate for the student's level of proficiency. In order to tailor the test closely to student ability levels while still including items and tasks that assess all of 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.

2.1.1. 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 Listening Subtest

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	12	PL1 - PL3	100%	0%	0%		
1	B	18	PL2 - PL4	78%	11%	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	95%	0%	5%		
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	100%	0%	0%	Dichotomous Selected Response	Machine Scored
6-8	A	12	PL1 - PL3	92%	0%	8%		
6-8	B	18	PL2 - PL4	55%	17%	28%		
6-8	C	18	PL3 - PL5	95%	0%	5%		
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 2.**Error! Reference source not found.** 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 2); 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 2); all others are routed into the B folder. Throughout the test, a student's underlying measure of ability is re-estimated with the completion of each folder, and the level of the next folder to be administered is chosen 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. 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 provided to test administrators in the Test Administrator Manual, the Listening subtest is untimed.

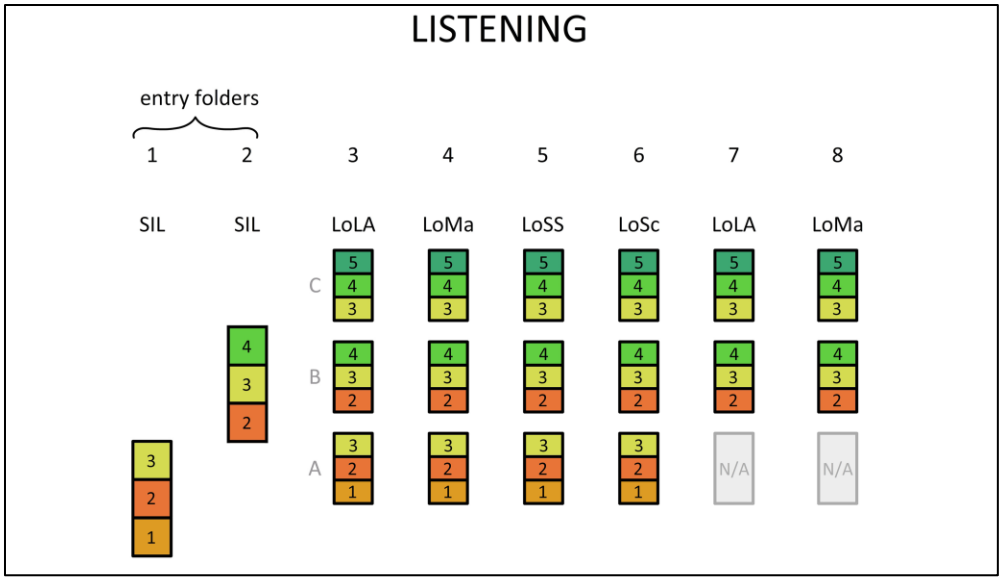


Figure 2. Format of the Listening test.

2.1.2. 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 Reading Subtest

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	92%	4%	4%		
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%	5%	0%		
4-5	B	24	PL2 - PL4	96%	4%	0%		
4-5	C	24	PL3 - PL5	96%	4%	0%		
6-8	Entry	6	PL1 - PL4	100%	0%	0%	Dichotomous Selected Response	Machine Scored
6-8	A	18	PL1 - PL3	90%	5%	5%		
6-8	B	24	PL2 - PL4	100%	0%	0%		
6-8	C	24	PL3 - PL5	96%	0%	4%		
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 3 shows the format of the Reading test. The format and adaptivity are similar to 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 this point. Although timing guidance is provided to test administrators in the Test Administrator Manual, the Reading subtest is untimed.

Figure 3. Format of the Reading test.

2.1.3. 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 Subtest

Grade-Level Cluster	Tier	Number of Tasks	*Targeted PL range*	Task Type	Response Formats	Scoring Procedures
1	A	2	PL1 - PL3	Writing Constructed Response	Polytomous Constructed Response; handwritten in test booklet	Human Scored: Centrally scored by DRC
1	B/C	2	PL2 - PL5			
2-3	A	2	PL1 - PL3	Writing Constructed Response	Polytomous Constructed Response; handwritten in test booklet	Human Scored: Centrally scored by DRC
2-3	B/C	2	PL2 - PL5			
4-5	A	2	PL1 - PL3	Writing Constructed Response	Polytomous Constructed Response; handwritten in response booklet or keyboarded in test platform	Human Scored: Centrally scored by DRC
4-5	B/C	2	PL2 - PL5			
6-8	A	2	PL1 - PL3	Writing Constructed Response	Polytomous Constructed Response; handwritten in response booklet or keyboarded in test platform	Human Scored: Centrally scored by DRC
6-8	B/C	2	PL2 - PL5			
9-12	A	2	PL1 - PL3	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 4, 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, all tasks are scored 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.

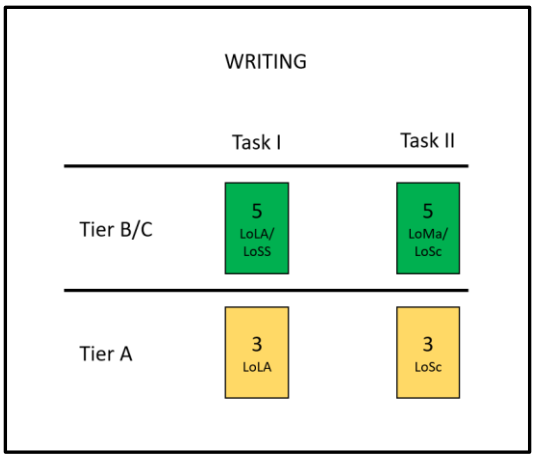


Figure 4. Format of the Writing test.

Beginning with Series 501, both tiers consist of two tasks. Prior to Series 501, all test forms had three tasks, with the exception of 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.[1] The ways in which the Standards are targeted by these tasks vary across grade levels and are spelled out in the generative item specifications.

[1] There are two exceptions to the distribution of the WIDA Standards on the Series 501 Writing subtest. For Grade 1, Tier A, Task II is written to the Social and Instructional WIDA Standard. This is due to the design of the embedded field test for items developed for Series 502, as described in Section 2.2.3 below. 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. It is anticipated that this slot will be refreshed in Series 503.

The design of the Writing field test for Series 501 is described in greater detail in Section 2.2.2.3 below.

Placement into tiers on the Writing test depends on how students perform on the Listening and Reading tests, which receive computerized scores. To determine how to best place students into a tier, test data for all students who were administered the assessment in the 2015–2016 operational year (the first year of the ACCESS Online assessment) were analyzed to examine the relationship between how students perform on Listening and Reading and how they perform on Writing, using logistic regression analyses. This information was used to program an algorithm into the ACCESS Online test that is used by the computer 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, based on their performances in Listening and Reading, into Tier B/C for Writing. All other students are placed into Tier A.

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

2.1.4. 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 Subtest

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

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

As shown in Figure 5, 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 a given 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.

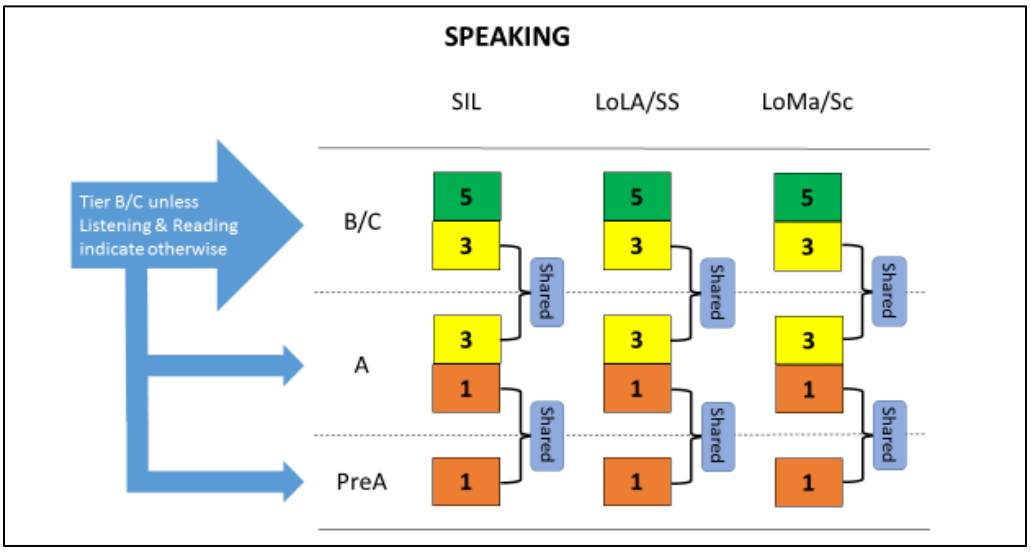


Figure 5. Format of the Speaking test.

As with Writing, placement into the three tiers on the Speaking test depends on 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 Listening and Reading 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 above, these tasks are targeted to the P1 level. These tasks are scored using a modified version of the full Speaking rating scale (see Section 3.2.4).

Placement of students into Tiers A and B/C in Writing is analogous to tier placement for Speaking. Test data for all students who were administered the assessment in the 2015–2016 operational year (the first year of the ACCESS Online assessment) were analyzed to examine the relationship between students' performance on Listening and Reading and performance on Speaking, using logistic regression analyses. This information was used 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, based on their performances in Listening and Reading, into Tier B/C for Speaking, and to place all other students into Tier A (except for those students, as noted above, who are routed into Tier Pre-A).

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

2.2. Test Construction

2.2.1. 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 Refreshment Plan is developed by the CAL Test Development (TD) team, and takes a number of 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 in the previous cycle. The Refreshment Plan is presented to WIDA for approval.

Upon receiving sign-off on the Refreshment Plan, CAL TD then determines which item specifications need to be updated or replaced and which can move forward as is. Generally, CAL TD updates or replaces item specifications for two reasons. On one hand, CAL TD analyzes prior items that did not perform as intended to determine if the poor performance was due to item mechanics or if a deeper item-specification issue was at fault. In the latter case, the specification can be updated (usually focused on updating the MPIs) or completely replaced, depending on the specific situation. On the other hand, CAL also updates or replaces item specifications as content standards change. As noted above, the ACCESS item specifications include explicit connections to the content standards. Should an update to the relevant content standard make an ACCESS item specification obsolete, CAL TD revises or replaces the specification.

Once updates to item specifications are complete, item development begins. The generation of raw item content occurs in two interconnected steps. First, CAL conducts what is called Theme Generation. In the ACCESS item specifications, each specification is written to a broad Topic related to the given WIDA 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 with one or more of the WIDA Standards, and these educators are provided with key parts of the item specification document, namely the Topic, the MPIs, and guidelines for selecting a good Theme. Then, CAL conducts brainstorming sessions via teleconference where the educators propose themes related to the topic that are grade-level appropriate. After the Theme Generation process is complete, CAL Language Testing Specialists and TD managers review the lists of themes to determine which move on to item writing. This determination is based on several factors, including operationalizability on a large-scale assessment, current themes on the assessment, and bias and sensitivity considerations.

Themes are then assigned 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 are also considered. All item writers, both new item writers and those returning from the previous cycle, receive an introductory training and are provided with extensive documentation regarding writing items for ACCESS, including an Item Writing Handbook and ancillary documents (checklists, item specifications, templates) to complete their assignments. Item writers are also assigned to work with one or more CAL Language Testing Specialists, who provide 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 will undergo further development and which will be retired. Folders then go to their first external review, Standards Expert review.

During 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 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 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 both peer -review by other Language Testing Specialists and 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 proficiency level of the stimulus, the graphic support, the question stems and response options (for Listening and Reading), and the task prompts (for Speaking and Writing). CAL TD staff also conduct mock administrations. During this phase, other ancillary materials, such as scripts and directions, are produced. Upon sign-off, TD staff work 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 contains all of the item content, including assets such as graphics and audio files, coded to be readable by the test engine. There is one QTI package for each folder on ACCESS. Once the graphics have been generated, they are inserted into the folders, and layout review and fact checking are conducted (with 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 conducted by educators and WIDA staff on ACCESS items. Items are submitted to the content review panel to ensure that the content is accessible and relevant to students in the targeted grade-level cluster and at the targeted proficiency level, and that each item or task matches the MPI from the WIDA ELD 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 panelists are educators with culturally and linguistically diverse backgrounds who have experience interacting with English learners from a range of cultural, regional, religious, linguistic, ethnic, and socioeconomic backgrounds. WIDA recruits educators with culturally and linguistically diverse backgrounds from WIDA Consortium states to participate in the review, and 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 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 manager sign-off as a final step. Graphics and the QTI packages are subsequently revised accordingly.

Tasks in the domain of Writing undergo one additional step: a small-scale tryout with educators and students. Given the changes to the Writing subtest 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 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 uses the student responses and the survey data in a qualitative analysis 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 2.2.2.3 for more information regarding the field test design.)

After Content Review and Bias and Sensitivity Review edits (and Tryouts edits for Writing) have been implemented, the folders are then prepared for final production steps. Test developers produce audio recording scripts for professional audio recording, arrange for recording the audio files, conduct final layout reviews, and perform key checks for Listening and Reading. Both CAL and WIDA conduct quality control checks of the QTI. WIDA signs off on all materials before Data Recognition Corporation (DRC) builds the final test forms in the test engine. Items that reach this point then go through field testing processes, described by domain below.

2.2.2. Field Testing

2.2.2.1. Listening

Listening items developed for Series 501 were field tested as embedded folders during the operational administration of Series 403. The embedded field test folders included innovative item formats, including hot spot items, where the student clicks on an area of the screen, and drag-and-drop items, where the student drags an image/text to a specified screen area to respond.

For Series 501, a total of 108 Listening items (36 folders) were field tested, across all five grade-level clusters, as indicated in Table 5.

Each student received one Listening field test folder embedded into 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, Standard, and tier pool target (Entry, A, B, or C) of the folder. Students are administered the embedded field test folder at the stage targeted for refreshment, with administration randomized so that half of the students see the field test folder before the corresponding operational folder, and half see the operational folder before the field test folder. Field test folders are administered to those students who are routed to take the operational folder that is either at the same tier or adjacent to the tier that the field test folder targets. When field test samples are drawn, the sample includes 50% of students at the tier targeted by the field test folder and 50% at adjacent tiers (if there are adjacent tiers both above and below, 25% from each). In cases where the folder to be field tested is to be placed in one of the entry stages, students who receive that field test folder will receive it directly after the pair of operational entry folders. Field test sample targets in Listening are set at a minimum of 3,000 responses per folder. Because the Listening field test data are also used in the pre-equating analysis, the sample size requirement of 3,000 is well in excess of the minimum of 250 per form for high stakes tests proposed by Linacre (1994), in order to ensure that the pre-equated parameter estimates are stable. Linacre (1994), citing Wright and Douglas's (1975) formulation, illustrated how to determine the minimum sample required for calibrating dichotomous-scored items to achieve various levels of estimation precision and confidence intervals. With a sample size of 3,000, we

can be 95% confident that no item parameter is more than plus or minus 0.1 logit away from its true values.

Table 5
Number of Series 501 Listening Field Test Folders and Items

Grade-Level Cluster	Tier Pool	Number Folders to refresh	Number overage folders	Total number of field test folders	Total number of field test items	Standards included in FT
1	Entry	0	0	0	0	
1	A	1	1	2	6	LoSS
1	B	1	1	2	6	LoSS
1	C	1	1	2	6	LoSS
2-3	Entry	1	1	2	6	SIL
2-3	A	1	1	2	6	LoLA
2-3	B	0	0	0	0	
2-3	C	1	1	2	6	LoLA
4-5	Entry	1	1	2	6	SIL
4-5	A	0	0	0	0	
4-5	B	1	1	2	6	LoMA
4-5	C	1	1	2	6	LoMA
6-8	Entry	1	1	2	6	SIL
6-8	A	1	1	2	6	LoSC
6-8	B	1	1	2	6	LoSC
6-8	C	1	1	2	6	LoSC
9-12	Entry	1	1	2	6	SIL
9-12	A	1	1	2	6	LoSC
9-12	B	1	1	2	6	LoSC
9-12	C	2	2	4	12	LoMA, LoSC
Total		**18**	**18**	**36**	**108**	

After field test data are drawn, folders of items are analyzed for their psychometric properties, and those that meet established psychometric standards are eligible for selection in the next year's operational test.

2.2.2.2. Reading

Reading items developed for Series 501 were field tested as embedded items during the operational administration of Series 403. All embedded field test items for Reading were traditional multiple-choice items. No innovative item formats were included in the Series 501 Reading field test.

For Series 501, a total of 192 Reading items (64 folders) were field tested, across all five grade-level clusters, as indicated in Table 6.

Table 6
Number of Series 501 Reading Field Test Folders and Items

Grade-Level Cluster	Tier Pool	Number Folders to refresh	Number overage folders	Total number of field test folders	Total number of field test items	Standards included in FT
1	Entry	0	0	0	0	
1	A	1	1	2	6	LoSS
1	B	2	2	4	12	LoLA, LoMA
1	C	3	3	6	18	LoSS, LoLA, LoMA
2-3	Entry	0	0	0	0	
2-3	A	2	2	4	12	LoSC, LoMA
2-3	B	2	2	4	12	LoSC, LoMA
2-3	C	2	2	4	12	LoSC, LoMA
4-5	Entry	1	1	2	6	SIL
4-5	A	2	2	4	12	LoLA, LoMA
4-5	B	2	2	4	12	LoLA, LoMA
4-5	C	2	2	4	12	LoLA, LoMA
6-8	Entry	1	1	2	6	SIL
6-8	A	1	1	2	6	LoLA
6-8	B	2	2	4	12	LoLA, LoMA
6-8	C	3	3	6	18	LoLA, LoMA
9-12	Entry	0	0	0	0	
9-12	A	2	2	4	12	LoMA, LoSS
9-12	B	2	2	4	12	LoMA, LoSS
9-12	C	2	2	4	12	LoMA, LoSS
Total		32	32	64	192	

The embedded Reading field test is administered in the same way as the embedded Listening field test. As with Listening, field test sample targets in Reading are set at a minimum of 3,000 responses per folder.

After field test data are drawn, folders of items are analyzed for their psychometric properties, and those that meet established psychometric standards are eligible for selection in the next year's operational test.

2.2.2.3. Writing

Series 501 Writing tasks were field tested in a small-scale stand-alone field test. For Series 501, a total of 15 Writing tasks were field tested, as indicated in Table 7.

Table 7
Number of Series 501 Writing Field Test Tasks

Grade-Level Cluster	Tier	Number of folders to refresh	Number of folders field tested	Standards included in FT
1	A	1	2	LoLA
1	BC	1	1	LoLA/LoSS
23	A	1	1	LoLA
23	BC	1	2	LoLA/LoSS
45	A	1	1	LoLA
45	BC	1	2	LoLA/LoSS
68	A	1	1	LoLA
68	BC	1	2	LoLA/LoSS
91	A	1	1	LoLA
91	BC	1	2	LoLA/LoSS
Total		**10**	**15**	

A sample of 500 students per task was targeted. This is well in excess of the minimum of 250 per form for high stakes tests proposed by Linacre (1994), and allows for at least 10 observations per category, as recommended by Linacre (2002) for polytomous items. Since the score distribution for Writing 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 in order to ensure that there will be students at the high end of the score distribution for analysis, as well as to ensure that students' Writing samples are available at those score points in order to create scoring materials.

The field test was administered 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 field test samples. DRC performed a 20% read-behind as a quality control measure, with the first score as the score of record.

Quantitative and qualitative analyses of the collected responses were conducted. The main purposes of this small-scale field testing were (a) to confirm that the tasks are working as intended, (b) to identify anchor samples for rater training, and (c) to inform the rating of the tasks when they become operational. Note that for the stand-alone Series 501 Writing field test, the sample target was not met for all clusters and tiers. Despite not meeting the sample targets, there were sufficient responses to conduct qualitative analyses, review raw score distributions, and provide evidence for the suitability of tasks for operational testing.

2.2.2.4. Speaking

All Tier A and B/C students are administered a Speaking field test folder appended to their operational Speaking assessment. Tier Pre-A is not included in the field test. A total of 54 tasks (18 panels) were field tested for Series 501, with a target sample size of 500 students per folder. This is well in excess of the minimum of 250 per form for high stakes tests proposed by Linacre (1994), and allows for at least 10 observations per category, as recommended by Linacre (2002) 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 in order 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 in order to create scoring materials.

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

Students receive a Speaking field test folder in the tier that corresponds to their operational tier. For Series 501, a total of 36 Speaking tasks were field tested, as indicated in Table 8.

Table 8
Number of Series 501 Speaking Field Test Tasks

Grade-Level Cluster	Tier	Number of folders to refresh	Number of folders field tested	Standards included in FT
1	A	2	4	SIL, LoMA/LoSC
1	BC	2	4	SIL, LoMA/LoSC
23	A	1	2	SIL
23	BC	1	2	SIL
45	A	2	4	SIL, LoMA/LoSC
45	BC	2	4	SIL, LoMA/LoSC
68	A	2	4	SIL, LoMA/LoSC
68	BC	2	4	SIL, LoMA/LoSC
91	A	2	4	SIL, LoMA/LoSC
91	BC	2	4	SIL, LoMA/LoSC
Total		**18**	**36**	

2.2.3. Item Selection

Subsequent to 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 501 operational assessment. Qualitative and quantitative methods guide the selection of operational items.

In the domains of Listening and Reading, item selection is a two-step process. First, the item selection panel reviewed the field test results. We use a three-tier color-coding system for field test review. Items are coded as "green," "yellow," or "red," and a folder is then colored based on the least favorable item in the folder. In other words, a folder with a red item is always coded as red, a folder with a yellow item (but no red items) is coded yellow, and folders are coded green only when all items are green.

Items are coded by color according to the following criteria:

- If an item shows C-level or CC-level differential item functioning (DIF), it is automatically coded yellow. Any items that show this level of DIF are subject to an extra round of review prior to item selection (see Part 2 Section 2.2 for further detail), and the item selection panel is provided with the report of the DIF review.
- Items are coded as green if they have infit and outfit values less than or equal to 1.20. As very easy items are particularly sensitive to outliers, any item with a *p* -value greater than 0.85 is automatically coded as green, even if it has fit values outside of these thresholds.
- Items with infit and outfit values greater than 1.20 and less than 1.50 are coded as yellow. As difficult items are also sensitive to outliers, items with *p* -values close to chance (0.40 for a 3-response item, and 0.35 for a 4-response item) are coded as yellow if outfit is greater than 1.20 and less than 1.75.
- Items that do not meet these criteria are coded as red.

The task of the item selection panel in this first stage is to review all yellow folders and recode them as "green," meaning "appropriate for operational use," or "red," meaning "not appropriate for operational use."

In the next stage, the set of green folders, which the panel has deemed appropriate for operational use, becomes the pool of folders for item selection. Folders are selected with attention to the difficulty of each item within a folder, the mean item difficulty of a folder, and the content of a folder.

Tables 9 and 10 provide numbers of continuing and new items per grade-level cluster for Listening and Reading. 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 9
Number of New and Continuing Items on ACCESS Online Series 501 Listening, 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 10
Number of New and Continuing Items on ACCESS Online Series 403 Reading, 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 considers both qualitative and quantitative analyses of the tasks. Test development specialists review student responses and DRC raters' comments on field-tested tasks. These observations are integrated with item statistics, including fit statistics, raw score distributions, and rater agreement, to produce a recommendation for the panel. The panel then reviews the recommendation and associated evidence and either accepts or rejects the recommendation.

Tables 11 and 12 provide numbers of continuing and new items, per grade-level cluster, for Writing and Speaking. 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 11
Number of New and Continuing Items on ACCESS Online Series 501 Writing, 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
	B/C	1	1	2
2–3	A	1	1	2
	B/C	1	1	2
4–5	A	1	1	2
	B/C	1	1	2
6–8	A	1	1	2
	B/C	0	2	2
9–12	A	1	1	2
	B/C	1	1	2

Table 12
Number of New and Continuing Tasks on ACCESS Online Series 501 Speaking, by Grade-Level Cluster

Grade-level cluster	Tier	Number of new tasks	Number of continuing tasks	Total number of tasks
1	Pre-A	2	1	3
1	A	4	2	6
1	B/C	4	2	6
2–3	Pre-A	2	1	3
2–3	A	4	2	6
2–3	B/C	4	2	6
4–5	Pre-A	2	1	3
4–5	A	4	2	6
4–5	B/C	4	2	6
6–8	Pre-A	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

2.3. Item and Task Design

This section describes how items and tasks are designed in order 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.

2.3.1. 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 illustrations; for Grades 2–12, items that test listening proficiency at PLs 3–5 may consist of short written text response options that are written to be about two PLs lower than the targeted PL of the Listening item. Most items on the operational Listening assessment 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, where the student clicks on an area of the screen, and drag-and-drop items, where the student drags an image/text to a specified screen area to respond. The number of enhanced items on the Listening subtest 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 is written to target the language of one of the five WIDA ELD Standards and to test a student's ability to process language at one of the five fully delineated proficiency levels. *Folders* group together three test items that are written around a common theme, with each item targeting a progressively higher proficiency level.

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

Listening items are developed so that each 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 produced by a professional recording studio. 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 3.4.2.1.

2.3.2. 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 illustrations rather than text. As with Listening items, Reading items are grouped into thematic folders of three test items each.

- 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 ACCESS Online Reading, 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 assessment 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, where 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 subtest 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 grade-level cluster and 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.

2.3.3. 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. As described in Section 2.1.3. above, these tasks are scored using the entire breadth of the scoring scale; therefore, students may achieve proficiency levels higher than PL 3, although the tasks are not designed to elicit extended responses, so the scores are limited by task design. Tasks appearing on the Tier B/C form are designed to 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, they are scored on the entire breadth of the scoring scale, so students' actual performances 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.

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.

2.3.4. 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 in order 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, test takers first listen to the model student respond to a parallel task. The purpose of the model is to demonstrate task expectations to both test takers 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 stimulus 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.[2] 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 6 shows the generic screen layout of the Speaking test.

[2] During the piloting of the Speaking test design prior to ACCESS Online going 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.

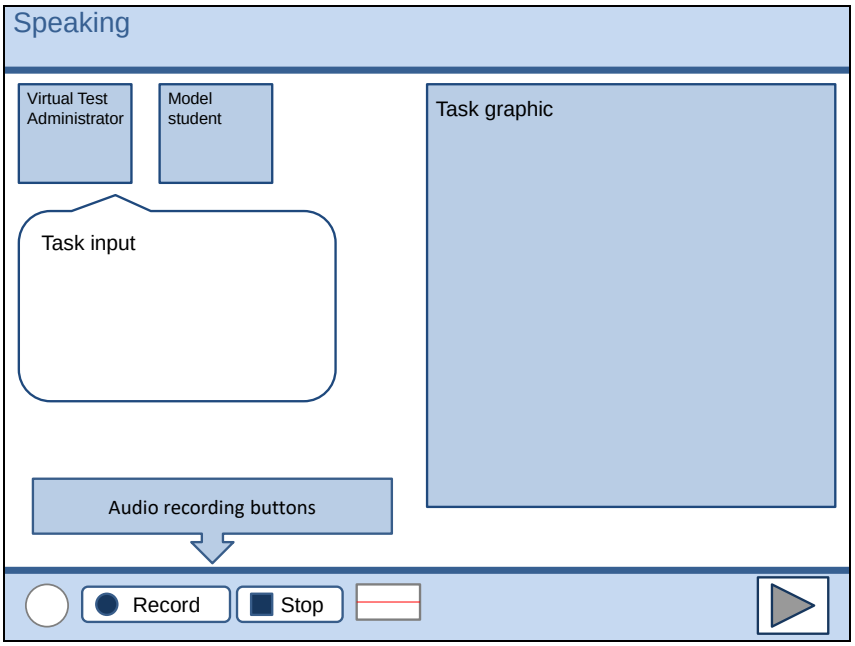


Figure 6. Visualization of the Speaking test screen layout.

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

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 5 in Section 2.1.4 above for an illustration).

3. Assessment Performance: The Implementation of ACCESS

3.1. Test Delivery

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

3.1.1. Listening and Reading

Listening and Reading are the first domains assessed. Students may take these in either order. Students sit at individual computer monitors and take the Listening and Reading tests online. They use headsets to listen to directions for the Listening and Reading tests, as well as to the Listening items. Students use the computer interface to select or record their answers; once a student records and 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 content 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 2.3.4).

3.2. Scoring Procedures

3.2.1. 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."

3.2.2. Scoring Performance-Based Tasks: Writing and Speaking

Performance-based tasks in the domains of Writing and Speaking are scored by trained raters. DRC retains a number of raters from year to year. This pool of experienced raters was drawn from to staff the scoring of the ACCESS for ELLs. To complete the rater staffing, recruiting events were held and applications for rater positions were screened by DRC's recruiting staff. Candidates were personally interviewed by DRC staff. In addition, each candidate was required 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, preference was given 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. These individuals were valued for their content-specific knowledge, but 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 in order to ensure that raters understand the nuances of each unique Writing or Speaking task. Team leaders, who are selected based on prior performance as raters and for their leadership skills, are assigned to small groups of raters; there are typically 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. Scoring directors are promoted from within DRC and earn their positions by demonstrating quality work as raters and as team leaders on previous projects. 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

- Raters are seated at stations and are assigned unique ID numbers and passwords.
- The scoring director provides detailed directions for use of DRC's computerized scoring system.
- 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, they are further trained for their grade-level cluster 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 are explained below) to keep the raters calibrated on the actual tasks they are scoring.

Calculating Score Agreement for Score Monitoring

- For Writing, agreement is defined as two adjacent scores. (See Section 3.2.3 for a description of the Writing Scoring Scale.) For example, using the Writing Scoring Scale, scores of 2 and 2+ would be considered agreement, as would scores of 2 and 2 or scores of 2+ and 3. Scores of 2 and 3 on the Writing Scoring Scale would be considered adjacent, and scores of 2 and 3+ would be considered nonadjacent.
- For Speaking, agreement is defined as two scores that are exactly the same. (See Section 3.2.4 for a description of the Speaking Scoring Scale.)

Routing Responses to Ensure "Blind" Second Ratings

- The DRC scoring system ensures that responses are routed to qualified raters until the prescribed number of ratings is performed 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. At least 20% of the responses are independently scored by two raters for the purpose of monitoring interrater reliability. DRC monitors these data daily.
- 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 Writing, DRC used both recalibration sets and validity responses to monitor handscoring quality control. Recalibration sets and validity responses were developed in conjunction with DRC, CAL, and WIDA. 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). This pool of responses and their scores were reviewed and

approved by WIDA staff. DRC supervisors supplemented this pool of responses as needed by selecting additional responses; these responses and their scores were reviewed and approved by CAL and WIDA before use. For each of the first 5 days that raters scored a task, they took 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 Writing 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. The validity responses were seeded into operational scoring; the raters did not know which responses were operational and which were validity responses. Reports generated on a daily basis compared the scores given by each rater to the "true" score for each validity response. When a rater was working on a task, the validity responses were dealt 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.

- For Speaking, DRC uses recalibration sets, which were developed in the same manner as the recalibration sets for Writing. As with Writing, for each of the first 5 days that raters scored a task, they took one recalibration set of five responses to ensure that they were calibrated. The raters' performances on recalibration sets were used for monitoring and maintaining reliability. After these first 5 days, recalibration sets were used twice weekly to monitor scoring. The functionality for providing validity sets for Speaking is currently in development and should be available for 2021 scoring. For administrations prior to the 2021 scoring, additional recalibration was used to provide ongoing quality control checks of raters' performance.

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 should be assigned), 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 that could not be assigned 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 in order 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 scorers apply a nonscorable code (with the exception of 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, potential plagiarism, or potential teacher interference), DRC's imaging system gave scorers 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 raters 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 in order 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 in order to score the response.

Changes in Scoring Procedures due to the COVID-19 Pandemic

During the second half of March 2020, DRC pivoted from site-based scoring to remote scoring in order to continue handscoring operations in the safest manner. DRC's remote scoring was designed 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 training was delivered (online) and in the modes of communication used (web screen sharing, webcast, video chat, and chat). Scoring leaders were equipped 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 through the use of the tools that DRC utilized for remote scoring. These training sessions took place in late March and were completed by early April. Once supervisors and raters were trained on the remote scoring process, handscoring resumed 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. Raters were also required to read and sign nondisclosure agreements. During training and large-group discussions, emphasis was always given to what security means, the importance of maintaining security, and how this is accomplished. In the remote environment, these security reminders were given daily. Raters working remotely were required to work in a private environment away from other people (including family members). 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.** Content training for operational items was already completed while raters were onsite. Additionally, approximately half of the field test training and scoring was completed onsite. For the remainder of the field test, content training was provided remotely, and it remained an interactive, comprehensive, hands-on experience. For Writing field test 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 field test 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. All secure materials such as sources, anchors, training sets,

and/or qualifying sets were accessible for raters and supervisors in ScoreBoard, which does not permit anything to be downloaded or printed. Scorers were not permitted to download, print, or screenshot any confidential materials, including test items and student responses. 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.

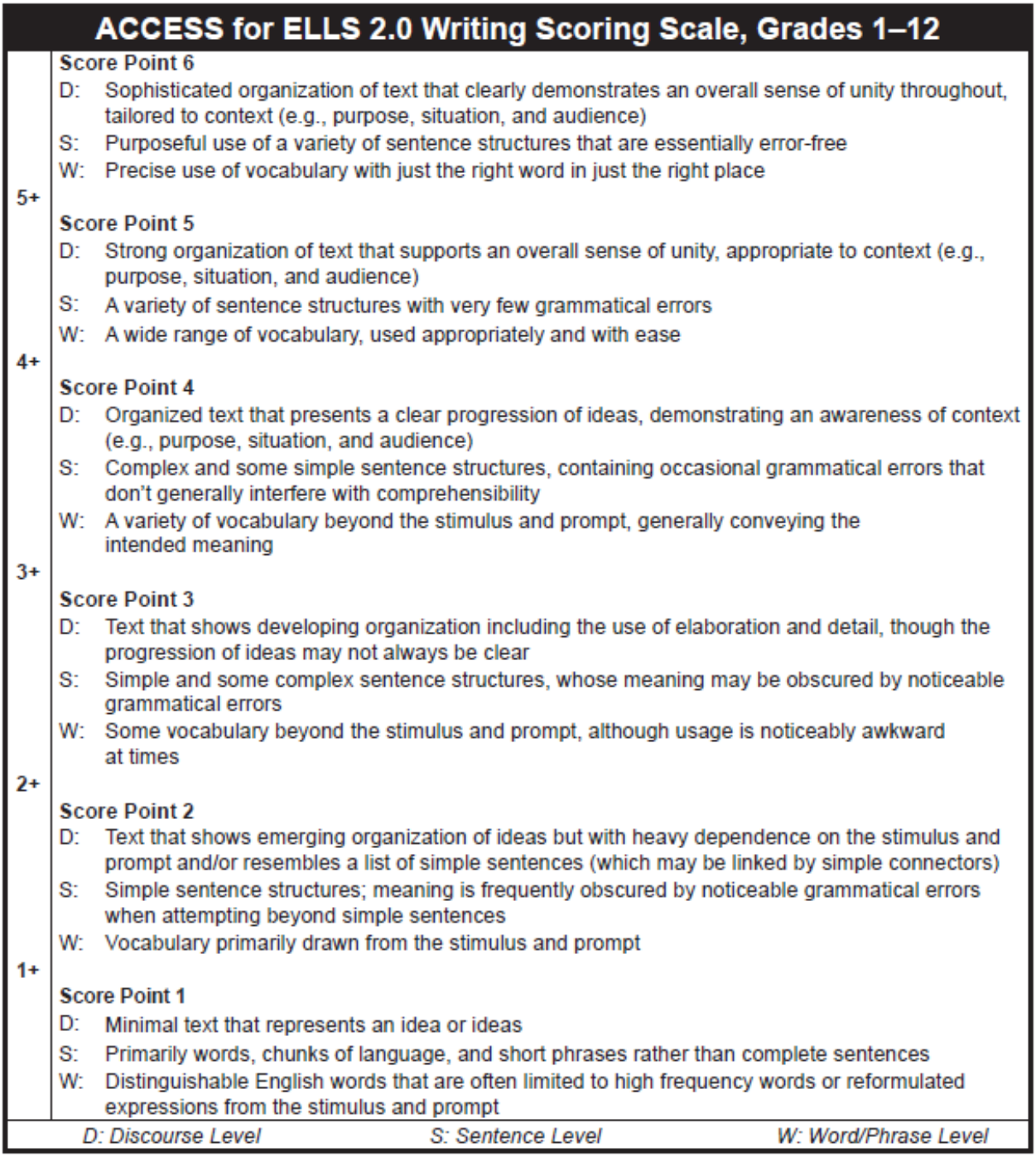
- **Quality control.** DRC's robust quality control processes and handscoring metrics were identical for onsite and remote scoring sessions. During remote scoring, scored responses were monitored with second reads exactly as they were at the scoring sites. Read-behinds were also conducted in the exact same manner; however, any conversations and/or retraining needed as a result of the monitoring were held in one-on-one video chat sessions. Handscoring quality reports continued to be available daily and on demand for handscoring supervisors and DRC's project leadership, and DRC continued to provide WIDA staffing with handscoring reports on the same schedule as when handscoring was onsite.

3.2.3. 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 in order 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 7 shows the Writing Scoring Scale.



ACCESS for ELLS 2.0 Writing Scoring Scale, Grades 1–12

	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
5+	
	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
4+	
	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
2+	
	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
1+	
	Score Point 1 D: Minimal text that represents an idea or ideas S: Primarily words, chunks of language, and short phrases rather than complete sentences W: Distinguishable English words that are often limited to high frequency words or reformulated expressions from the stimulus and prompt

D: Discourse Level *S: Sentence Level* *W: Word/Phrase Level*

Figure 7. 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; or is entirely in a language other than English.
>
> **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 13. Raw scores for the two operational tasks are added, giving a total raw score that ranges from 0 to 18.

Table 13

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.

3.2.4. 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 2.3.4). 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 8 shows the Speaking Scoring Scale.

ACCESS for ELLs 2.0 Speaking Scoring Scale	
Score point	**Response characteristics**
Exemplary use of oral language to provide an elaborated response	• Language use comparable to or going beyond the model in sophistication • Clear, automatic, and fluent delivery • Precise and appropriate word choice
Strong use of oral language to provide a detailed response	• Language use approaching that of model in sophistication, though not as rich • Clear delivery • Appropriate word choice
Adequate use of oral language to provide a satisfactory response	• Language use not as sophisticated as that of model • Generally comprehensible use of oral language • Adequate word choice
Attempted use of oral language to provide a response in English	• Language use does not support an adequate response • Comprehensibility may be compromised • Word choice may not be fully adequate
No response (in English)	• Does not respond (in English)

Figure 8. 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 14. 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 14

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.

3.3. Operational Administration

3.3.1. 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.

3.3.2. 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.

3.3.3. 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.

3.3.3.1. 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 are able to 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 are able to 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 are able to 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.

3.3.4. 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.

3.3.4.1. 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.

3.3.5. Speaking Test Administration

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

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.

3.3.5.1. 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 are able to 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. The majority of students are placed in either Tier A or Tier B/C.

3.3.5.2. 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 than 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.

3.3.6. 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.

3.4. Accessibility and Fairness

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.)

3.4.1. Support Provided to All ELLs

Universal design. ACCESS for ELLs incorporates universal design principles in order 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 in order 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.

3.4.2. 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 comparable test results 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 test developers at the Center for Applied Linguistics.

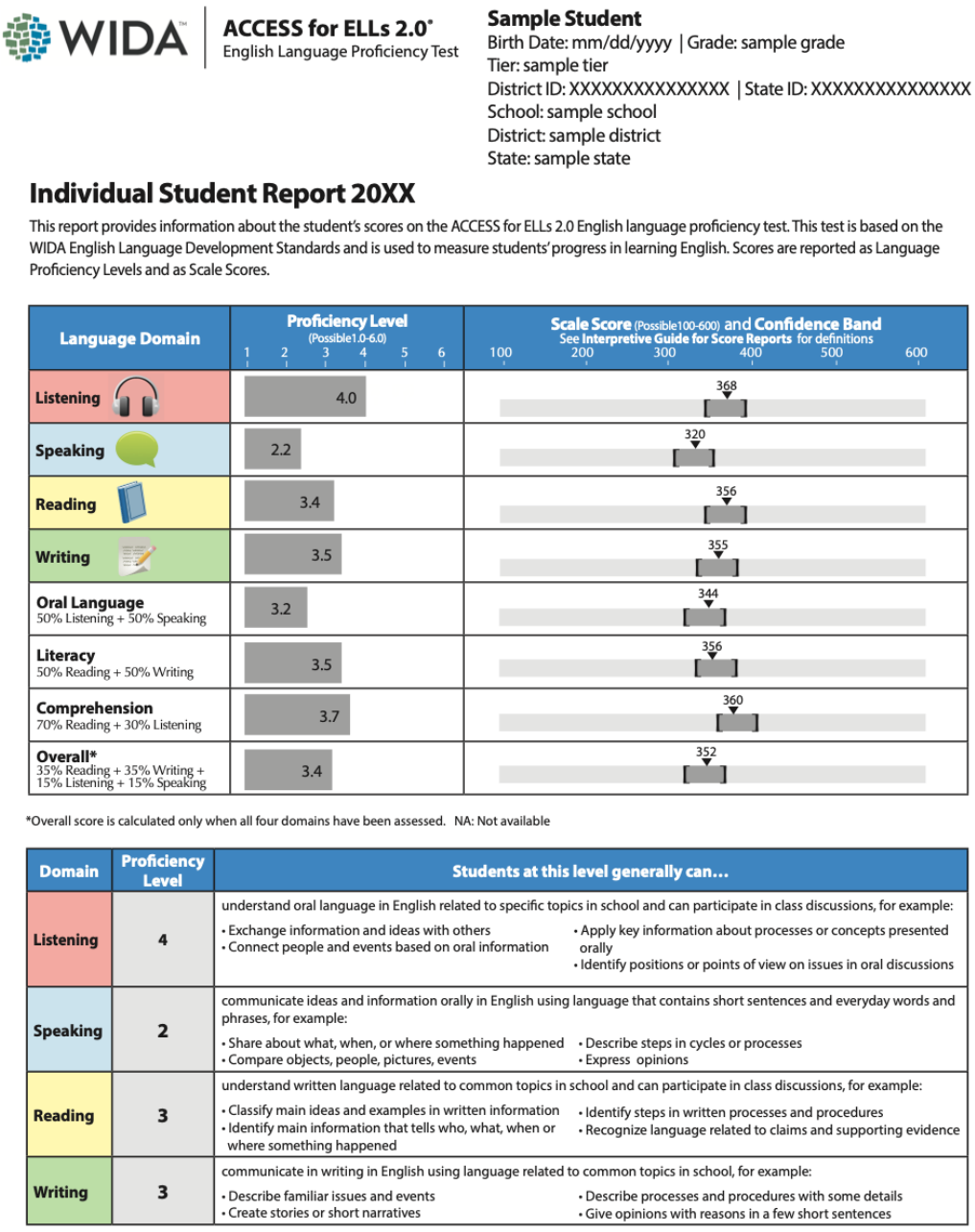
WIDA offers *Alternate ACCESS for ELLs.* This test is intended only for those ELLs who have cognitive disabilities that are so significant as to prevent meaningful participation in ACCESS testing, even with accommodations. The results of the Alternate ACCESS for ELLs operational administration appear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

4.1. Individual Student Report

The Individual Student Report (Figure 9) 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 one indicator of a student's English language proficiency, the language needed to access content and succeed in school.



WIDA | ACCESS for ELLs 2.0® English Language Proficiency Test

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

Individual Student Report 20XX

This report provides information about the student's scores on the ACCESS for ELLs 2.0 English language proficiency test. This test is based on the WIDA English Language Development Standards and is used to measure students' progress in learning English. Scores are reported as Language Proficiency Levels and as Scale Scores.

Language Domain	Proficiency Level (Possible1.0-6.0)	Scale Score (Possible100-600) 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: • Exchange information and ideas with others • Connect people and events based on oral information • Apply key information about processes or concepts presented orally • Identify positions or points of view on issues in oral discussions
Speaking	2	communicate ideas and information orally in English using language that contains short sentences and everyday words and phrases, for example: • Share about what, when, or where something happened • Compare objects, people, pictures, events • Describe steps in cycles or processes • Express opinions
Reading	3	understand written language related to common topics in school and can participate in class discussions, for example: • Classify main ideas and examples in written information • Identify main information that tells who, what, when or where something happened • Identify steps in written processes and procedures • Recognize language related to claims and supporting evidence
Writing	3	communicate in writing in English using language related to common topics in school, for example: • Describe familiar issues and events • Create stories or short narratives • Describe processes and procedures with some details • Give opinions with reasons in a few short sentences

Figure 9. 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.

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:

- 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; a particularly high score in one domain may effectively raise a low score in another. Similar overall scores can mask very different performances on the 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 scale, so scale scores should not be compared, such as comparing Listening to Reading. Proficiency level scores can be used for such comparisons.
- 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.

4.2. 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 Student Participation and Performance

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 501 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 were included in the descriptions of student participation and performance but were excluded from subsequent analyses, namely, 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 (i.e., more than 1) 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, 5% to 6% of students were excluded in each cluster and domain.

In addition to these data exclusions, 624 student records were removed from the data set due to a concern over plagiarized responses on a 9–12 Tier B/C Speaking task. Further detail on this issue can be found in WIDA's 2019–2020 *Year in Review Report*.

Table 1.1
Students Excluded from Analysis Due to Test Interruptions by Domain and Cluster

Domain	Cluster	No. of Students	Total Students	Percent
Listening	1	10,249	186,970	5.48%
	2–3	19,592	386,381	5.07%
	4–5	18,276	334,213	5.47%
	6–8	24,864	331,917	7.49%
	9–12	22,390	332,408	6.74%
	Total	95,371	1,571,889	6.07%
Reading	1	7,033	186,970	3.76%
	2–3	19,425	386,381	5.03%
	4–5	24,212	334,213	7.24%
	6–8	27,337	331,917	8.24%
	9–12	26,972	332,408	8.11%
	Total	104,979	1,571,889	6.68%
Writing	1	n/a	186,970	n/a
	2–3	n/a	386,381	n/a
	4–5	15,652	334,213	4.68%
	6–8	19,463	331,917	5.86%
	9–12	14,040	332,408	4.22%
	Total	49,155	998,538	4.92%
Speaking	1	11,958	186,970	6.40%
	2–3	22,171	386,381	5.74%
	4–5	23,241	334,213	6.95%
	6–8	24,055	331,917	7.25%
	9–12	22,354	332,408	6.72%
	Total	103,779	1,571,889	6.60%

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 2019–2020 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 states with more than 100,000 students were Illinois, North Carolina, and Georgia. The state with the smallest number of participants was the U.S. Virgin Islands. The biggest cluster was Grades

2–3. The territory abbreviations are as follows: DC, District of Columbia, MP, Northern Mariana Islands; and VI, U.S. Virgin Islands. BI indicates Bureau of Indian Education.

Table 1.1.1.1

Participation by Cluster by State, S501 Online

State	Cluster					Total
	1	2–3	4–5	6–8	9–12	
AK	1,061	2,332	2,543	3,009	2,632	11,577
AL	3,563	7,567	6,933	5,858	4,524	28,445
BI	213	480	521	813	349	2,376
CO	9,583	19,678	16,428	17,884	17,291	80,864
DC	534	1,122	1,117	744	950	4,467
DE	1,501	3,304	3,026	2,505	2,128	12,464
GA	14,302	28,416	25,601	20,995	18,363	107,677
HI	1,992	3,965	3,399	3,479	3,185	16,020
ID	2,156	4,655	4,338	4,245	3,501	18,895
IL	23,818	50,474	44,789	42,386	32,985	194,452
IN	7,655	16,022	14,381	12,340	13,533	63,931
KY	4,259	7,134	5,568	4,879	6,074	27,914
MA	11,574	21,381	15,042	16,317	19,993	84,307
MD	10,643	21,365	16,351	14,962	18,457	81,778
ME	517	962	881	906	1,090	4,356
MI	7,970	17,501	15,262	18,329	21,694	80,756
MN	7,835	15,911	12,561	12,000	12,923	61,230
MO	4,083	7,958	6,148	5,937	5,615	29,741
MP	77	273	280	375	224	1,229
MT	340	605	727	930	481	3,083
NC	12,367	26,498	25,084	23,657	21,489	109,095
ND	378	860	713	753	882	3,586
NH	477	1,014	812	932	987	4,222
NJ	5,960	11,613	9,570	8,812	10,196	46,151
NM	4,298	9,422	10,510	12,494	11,325	48,049
NV	6,385	13,791	12,076	13,438	15,472	61,162
OK	6,419	12,996	11,158	10,654	9,136	50,363
PA	5,935	13,208	12,384	14,896	17,262	63,685
RI	1,355	2,871	2,824	3,117	3,912	14,079
SC	2,244	5,123	5,018	6,387	7,583	26,355
SD	704	1,368	1,146	1,141	1,056	5,415
TN	3,998	7,224	5,176	5,433	6,282	28,113
UT	4,884	11,133	11,586	11,994	8,735	48,332
VA	12,297	25,780	18,936	16,966	21,073	95,052
VI	3	18	41	57	39	158
VT	157	379	334	298	352	1,520
WI	5,162	11,339	10,473	11,560	10,076	48,610

WY	271	639	476	435	559	2,380
Total	186,970	386,381	334,213	331,917	332,408	1,571,889

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 46% female and 52% male in Clusters 1–3 and 44% female and 54% male for Clusters 4–12. About 64% of participants were Hispanic in all clusters.

Table 1.1.1.2

Participation by Cluster by Gender, S501 Online

Cluster		Gender			Total
		F	M	Missing	
1	Count	85,583	96,326	5,061	186,970
	% within Cluster	45.8%	51.5%	2.7%	100.0%
2–3	Count	175,942	199,909	10,530	386,381
	% within Cluster	45.5%	51.7%	2.7%	100.0%
4–5	Count	148,515	177,163	8,535	334,213
	% within Cluster	44.4%	53.0%	2.6%	100.0%
6–8	Count	140,669	181,829	9,419	331,917
	% within Cluster	42.4%	54.8%	2.8%	100.0%
9–12	Count	140,335	182,079	9,994	332,408
	% within Cluster	42.2%	54.8%	3.0%	100.0%
Total	Count	691,044	837,306	43,539	1,571,889
	% within Cluster	44.0%	53.3%	2.8%	100.0%

Table 1.1.1.3

Participation by Cluster by Ethnicity, S501 Online

Cluster		Hispanic/Non-Hispanic			Total
		Hispanic	Other	Unknown	
1	Count	116,697	57,315	12,958	186,970
	% within Cluster	62.40%	30.70%	6.90%	100.00%
2–3	Count	245,635	112,774	27,972	386,381
	% within Cluster	63.60%	29.20%	7.20%	100.00%
4–5	Count	221,216	83,915	29,082	334,213
	% within Cluster	66.20%	25.10%	8.70%	100.00%
6–8	Count	217,477	78,775	35,665	331,917
	% within Cluster	65.50%	23.70%	10.70%	100.00%
9–12	Count	208,806	85,772	37,830	332,408
	% within Cluster	62.80%	25.80%	11.40%	100.00%
Total	Count	1,009,831	418,551	143,507	1,571,889
	% within Cluster	64.20%	26.60%	9.10%	100.00%

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 2% to 6%.

Table 1.1.1.4

Participation by Cluster by Tier by Domain, S501 Online

Cluster			Domain	
			Writing	Speaking
1	Tier	Pre-A	-	7,325
		A	158,531	72,395
		BC	28,413	107,236
	Total		186,944	186,956
2–3	Tier	Pre-A	-	17,547
		A	95,709	86,790
		BC	290,603	282,032
	Total		386,312	386,369
4–5	Tier	Pre-A	-	6,542
		A	51,989	34,005
		BC	282,213	293,658
	Total		334,202	334,205
6–8	Tier	Pre-A	-	9,774
		A	115,872	66,245
		BC	216,031	255,883
	Total		331,903	331,902
9–12	Tier	Pre-A	-	20,389
		A	120,158	137,081
		BC	212,214	174,908
	Total		332,372	332,378

1.1.2 Grade

This section provides tables parallel to those in the previous section, but broken out by grade rather than by grade-level cluster. Table 1.1.2.1 shows student counts by grade and state. The largest grade was 2nd grade and the smallest was 12th grade. Table 1.1.2.4 presents the percentages between Tier A and B/C and indicates that 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, S501 Online

State	Grade												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
AK	1,061	1,145	1,187	1,296	1,247	1,135	1,016	858	788	669	643	532	11,577
AL	3,563	3,885	3,682	3,715	3,218	2,534	1,960	1,364	1,618	1,239	921	746	28,445
BI	213	227	253	294	227	290	249	274	76	80	105	88	2,376
CO	9,583	9,920	9,758	8,775	7,653	6,239	6,026	5,619	5,455	4,586	3,841	3,409	80,864
DC	534	539	583	610	507	261	311	172	305	216	209	220	4,467
DE	1,501	1,667	1,637	1,670	1,356	1,058	815	632	841	525	421	341	12,464
GA	14,302	14,283	14,133	14,446	11,155	8,093	6,955	5,947	7,157	4,903	3,449	2,854	107,677
HI	1,992	2,023	1,942	1,972	1,427	1,486	1,091	902	1,043	828	671	643	16,020
ID	2,156	2,228	2,427	2,490	1,848	1,667	1,412	1,166	1,193	952	702	654	18,895
IL	23,818	25,186	25,288	24,175	20,614	16,746	14,549	11,091	10,605	8,785	7,297	6,298	194,452
IN	7,655	8,075	7,947	7,813	6,568	5,148	3,979	3,213	3,318	3,522	3,966	2,727	63,931
KY	4,259	3,496	3,638	3,164	2,404	1,724	1,662	1,493	2,111	1,579	1,242	1,142	27,914
MA	11,574	11,270	10,111	8,677	6,365	5,443	5,541	5,333	7,182	5,301	4,184	3,326	84,307
MD	10,643	10,799	10,566	9,487	6,864	5,429	5,007	4,526	8,063	4,328	2,745	3,321	81,778
ME	517	488	474	461	420	300	309	297	292	288	249	261	4,356
MI	7,970	8,580	8,921	8,519	6,743	6,180	6,205	5,944	6,556	5,705	4,721	4,712	80,756
MN	7,835	8,187	7,724	7,212	5,349	4,234	3,916	3,850	4,107	3,447	2,859	2,510	61,230
MO	4,083	4,074	3,884	3,521	2,627	2,052	1,969	1,916	1,870	1,504	1,194	1,047	29,741
MP	77	121	152	150	130	119	126	130	71	65	45	43	1,229
MT	340	300	305	317	410	385	325	220	152	130	120	79	3,083
NC	12,367	13,269	13,229	13,355	11,729	10,008	7,601	6,048	7,338	5,626	4,773	3,752	109,095
ND	378	461	399	391	322	295	233	225	240	223	200	219	3,586
NH	477	530	484	479	333	301	345	286	334	243	207	203	4,222
NJ	5,960	5,973	5,640	5,641	3,929	3,048	2,983	2,781	3,347	2,744	2,266	1,839	46,151
NM	4,298	4,691	4,731	5,355	5,155	4,662	4,241	3,591	3,623	3,156	2,475	2,071	48,049
NV	6,385	6,632	7,159	6,841	5,235	4,318	4,577	4,543	4,490	4,081	3,665	3,236	61,162
OK	6,419	6,410	6,586	6,123	5,035	4,185	3,602	2,867	2,598	2,546	2,167	1,825	50,363
PA	5,935	6,524	6,684	6,575	5,809	5,241	4,856	4,799	5,216	4,564	3,892	3,590	63,685
RI	1,355	1,406	1,465	1,544	1,280	1,112	1,057	948	1,202	1,047	872	791	14,079
SC	2,244	2,450	2,673	2,862	2,156	2,162	2,160	2,065	2,604	1,904	1,638	1,437	26,355
SD	704	688	680	619	527	478	370	293	352	251	237	216	5,415
TN	3,998	3,795	3,429	3,085	2,091	1,824	1,868	1,741	2,462	1,692	1,188	940	28,113
UT	4,884	5,339	5,794	5,893	5,693	4,830	4,063	3,101	2,388	2,422	2,218	1,707	48,332
VA	12,297	13,350	12,430	11,260	7,676	6,092	5,762	5,112	8,069	5,266	4,460	3,278	95,052
VI	3	9	9	23	18	26	20	11	16	12	9	2	158
VT	157	197	182	194	140	114	101	83	89	93	75	95	1,520
WI	5,162	5,736	5,603	5,582	4,891	3,979	4,048	3,533	3,014	2,638	2,356	2,068	48,610
WY	271	308	331	296	180	137	162	136	181	115	116	147	2,380
Total	186,970	194,261	192,120	184,882	149,331	123,335	111,472	97,110	110,366	87,275	72,398	62,369	1,571,889

Table 1.1.2.2

Participation by Grade by Gender, S501 Online

Grade		Gender			Total
		F	M	Missing	
1	Count	85,583	96,326	5,061	186,970
	% within Grade	45.8%	51.5%	2.7%	100.0%
2	Count	88,813	100,170	5,278	194,261
	% within Grade	45.7%	51.6%	2.7%	100.0%
3	Count	87,129	99,739	5,252	192,120
	% within Grade	45.4%	51.9%	2.7%	100.0%
4	Count	82,956	97,536	4,390	184,882
	% within Grade	44.9%	52.8%	2.4%	100.0%
5	Count	65,559	79,627	4,145	149,331
	% within Grade	43.9%	53.3%	2.8%	100.0%
6	Count	52,615	67,133	3,587	123,335
	% within Grade	42.7%	54.4%	2.9%	100.0%
7	Count	46,984	61,311	3,177	111,472
	% within Grade	42.1%	55.0%	2.9%	100.0%
8	Count	41,070	53,385	2,655	97,110
	% within Grade	42.3%	55.0%	2.7%	100.0%
9	Count	45,349	61,891	3,126	110,366
	% within Grade	41.1%	56.1%	2.8%	100.0%
10	Count	36,473	48,177	2,625	87,275
	% within Grade	41.8%	55.2%	3.0%	100.0%
11	Count	31,191	38,756	2,451	72,398
	% within Grade	43.1%	53.5%	3.4%	100.0%
12	Count	27,322	33,255	1,792	62,369
	% within Grade	43.8%	53.3%	2.9%	100.0%
Total	Count	691,044	837,306	43,539	1,571,889
	% within Grade	44.0%	53.3%	2.8%	100.0%

Table 1.1.2.3

Participation by Grade by Ethnicity, S501 Online

Grade		Hispanic/Non-Hispanic			Total
		Hispanic	Other	Unknown	
1	Count	116,697	57,315	12,958	186,970
	% within Grade	62.4%	30.7%	6.9%	100.0%
2	Count	121,753	58,356	14,152	194,261
	% within Grade	62.7%	30.0%	7.3%	100.0%
3	Count	123,882	54,418	13,820	192,120
	% within Grade	64.5%	28.3%	7.2%	100.0%
4	Count	121,180	48,348	15,354	184,882
	% within Grade	65.5%	26.2%	8.3%	100.0%
5	Count	100,036	35,567	13,728	149,331
	% within Grade	67.0%	23.8%	9.2%	100.0%
6	Count	82,262	28,251	12,822	123,335
	% within Grade	66.7%	22.9%	10.4%	100.0%
7	Count	73,079	26,482	11,911	111,472
	% within Grade	65.6%	23.8%	10.7%	100.0%
8	Count	62,136	24,042	10,932	97,110
	% within Grade	64.0%	24.8%	11.3%	100.0%
9	Count	72,660	25,489	12,217	110,366
	% within Grade	65.8%	23.1%	11.1%	100.0%
10	Count	55,488	21,890	9,897	87,275
	% within Grade	63.6%	25.1%	11.3%	100.0%
11	Count	43,700	19,929	8,769	72,398
	% within Grade	60.4%	27.5%	12.1%	100.0%
12	Count	36,958	18,464	6,947	62,369
	% within Grade	59.3%	29.6%	11.1%	100.0%
Total	Count	1,009,831	418,551	143,507	1,571,889
	% within Grade	64.2%	26.6%	9.1%	100.0%

Table 1.1.2.4

Participation by Grade by Tier by Domain, S501 Online

Grade			Domain	
			Writing	Speaking
1	Tier	Pre-A	-	7,325
		A	158,531	72,395
		BC	28,413	107,236
	Total		186,944	186,956
2	Tier	Pre-A	-	7,449
		A	57,018	46,431
		BC	137,205	140,373
	Total		194,223	194,253
3	Tier	Pre-A	-	10,098
		A	38,691	40,359
		BC	153,398	141,659
	Total		192,089	192,116
4	Tier	Pre-A	-	2,371
		A	26,065	19,197
		BC	158,811	163,309
	Total		184,876	184,877
5	Tier	Pre-A	-	4,171
		A	25,924	14,808
		BC	123,402	130,349
	Total		149,326	149,328
6	Tier	Pre-A	-	2,316
		A	35,310	20,311
		BC	88,021	100,704
	Total		123,331	123,331
7	Tier	Pre-A	-	3,664
		A	41,327	17,167
		BC	70,139	90,636
	Total		111,466	111,467
8	Tier	Pre-A	-	3,794
		A	39,235	28,767
		BC	57,871	64,543
	Total		97,106	97,104

Grade			Domain	
			Writing	Speaking
9	Tier	Pre-A	-	6,399
		A	49,735	61,634
		BC	60,622	42,325
	Total		110,357	110,358
10	Tier	Pre-A	-	5,412
		A	31,259	34,555
		BC	56,008	47,301
	Total		87,267	87,268
11	Tier	Pre-A	-	4,202
		A	22,303	14,360
		BC	50,087	53,832
	Total		72,390	72,394
12	Tier	Pre-A	-	4,376
		A	16,861	26,532
		BC	45,497	31,450
	Total		62,358	62,358

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. 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, S501 Online

Cluster		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Mean	320.27	287.07	255.15	255.02	287.90	271.26	297.08	276.11
	N	176,572	179,739	186,850	174,883	165,935	179,697	170,589	160,535
2–3	Mean	337.93	325.84	303.84	275.39	306.92	314.96	329.50	312.37
	N	366,603	366,612	386,137	364,084	346,854	366,461	349,816	331,643
4–5	Mean	414.14	356.32	337.86	313.60	364.24	347.14	373.73	352.12
	N	315,715	309,547	318,325	310,791	295,137	297,848	295,028	267,689
6–8	Mean	399.67	350.55	323.78	315.15	357.80	337.21	365.54	343.32
	N	306,619	304,091	312,084	307,579	286,552	290,946	286,279	259,225
9–12	Mean	394.68	378.95	349.39	311.29	353.27	364.36	383.93	360.93
	N	309,545	304,775	317,941	309,552	290,272	294,810	288,077	264,160

Table 1.2.1.2

Mean Scale Scores by Gender, S501 Online

Cluster	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	F	Mean	323.41	289.28	260.08	261.20	292.54	274.81	299.53	279.93
		N	80,990	82,091	85,531	80,410	76,474	82,077	78,079	73,856
	M	Mean	317.58	285.34	251.08	249.95	284.03	268.37	295.11	272.98
		N	90,826	92,738	96,259	89,729	84,983	92,710	87,877	82,312
	Missing	Mean	318.10	282.85	249.24	246.20	282.19	266.28	293.42	270.76
		N	4,756	4,910	5,060	4,744	4,478	4,910	4,633	4,367
2–3	F	Mean	339.59	327.78	309.29	281.19	310.60	318.65	331.31	315.97
		N	167,269	166,352	175,830	166,289	158,708	166,290	159,056	151,255
	M	Mean	336.73	324.37	299.28	270.59	303.96	311.97	328.15	309.44
		N	189,385	190,181	199,786	187,756	178,635	190,101	181,195	171,238
	Missing	Mean	332.68	321.60	299.06	268.98	301.09	310.56	325.04	307.62
		N	9,949	10,079	10,521	10,039	9,511	10,070	9,565	9,150
4–5	F	Mean	413.53	358.26	343.27	317.05	365.59	350.83	374.89	355.04
		N	140,628	137,038	141,223	138,588	131,896	131,803	130,963	119,229
	M	Mean	414.72	354.78	333.50	310.91	363.24	344.18	372.83	349.79
		N	167,096	164,533	168,951	164,232	155,741	158,373	156,526	141,613
	Missing	Mean	412.70	354.59	334.56	309.09	361.19	344.66	372.15	349.36
		N	7,991	7,976	8,151	7,971	7,500	7,672	7,539	6,847
6–8	F	Mean	398.52	353.44	328.29	317.44	358.28	340.93	367.16	345.94
		N	130,761	128,769	131,943	130,328	122,185	123,163	122,045	110,594
	M	Mean	400.57	348.38	320.27	313.44	357.45	334.35	364.31	341.26
		N	167,193	166,679	171,207	168,432	156,218	159,456	156,169	141,229
	Missing	Mean	399.75	349.36	324.57	313.69	357.42	337.19	364.75	343.30
		N	8,665	8,643	8,934	8,819	8,149	8,327	8,065	7,402
9–12	F	Mean	394.60	382.19	354.41	314.21	354.63	368.52	386.13	364.17
		N	131,413	128,417	133,982	130,666	123,197	124,090	122,009	111,781
	M	Mean	394.54	376.41	345.59	309.30	352.23	361.18	382.13	358.41
		N	168,926	167,220	174,305	169,513	158,406	161,820	157,509	144,466
	Missing	Mean	398.24	379.74	348.26	306.45	353.03	364.31	385.74	361.17
		N	9,206	9,138	9,654	9,373	8,669	8,900	8,559	7,913

Table 1.2.1.3

Mean Scale Scores by Ethnicity, S501 Online

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Non-Hispanic Asian	Mean	332.64	303.85	271.60	263.85	298.47	287.94	312.60	291.04
		N	24,293	24,495	25,567	24,002	22,924	24,491	23,398	22,112
	Non-Hispanic Pacific Islander	Mean	304.45	281.21	254.89	247.47	275.55	268.08	287.94	269.73
		N	1,531	1,571	1,635	1,534	1,439	1,570	1,483	1,398
	Non-Hispanic Black	Mean	321.41	290.20	258.01	267.61	294.81	274.31	299.71	280.53
		N	9,520	9,727	10,173	9,462	8,893	9,724	9,147	8,560
	Hispanic (Of Any Race)	Mean	316.70	282.63	250.64	251.50	284.33	266.79	292.89	271.87
		N	110,275	112,556	116,637	109,355	103,764	112,530	106,788	100,602
	Non-Hispanic American Indian	Mean	320.34	280.81	250.12	250.65	285.91	265.64	292.53	271.63
		N	1,175	1,205	1,238	1,141	1,088	1,204	1,150	1,065
	Non-Hispanic Multiracial	Mean	335.62	299.11	266.53	266.48	301.64	283.10	310.07	288.51
		N	782	780	824	772	734	779	742	696
	Non-Hispanic White	Mean	331.18	293.59	263.81	264.89	298.36	278.81	304.94	284.54
		N	16,805	17,001	17,837	16,668	15,771	16,997	16,118	15,153
	Unknown	Mean	312.96	283.41	248.91	246.39	280.10	266.26	292.26	270.31
		N	12,191	12,404	12,939	11,949	11,322	12,402	11,763	10,949
2–3	Non-Hispanic Asian	Mean	353.87	338.18	317.61	283.84	319.08	328.15	342.96	325.23
		N	45,503	45,585	47,788	45,257	43,292	45,576	43,651	41,628
	Non-Hispanic Pacific Islander	Mean	321.45	319.24	303.18	263.21	292.73	311.34	320.05	305.60
		N	3,085	3,127	3,326	3,122	2,912	3,127	2,936	2,777
	Non-Hispanic Black	Mean	339.10	326.39	304.96	283.86	311.82	315.79	330.33	314.53
		N	20,094	20,174	21,427	20,062	18,905	20,171	19,053	17,971
	Hispanic (Of Any Race)	Mean	333.79	323.07	300.69	272.90	303.59	312.00	326.30	309.27
		N	233,559	233,634	245,497	231,733	221,083	233,529	223,201	211,646
	Non-Hispanic American Indian	Mean	335.79	319.14	299.19	266.83	301.49	309.29	324.17	306.65
		N	2,303	2,314	2,418	2,260	2,136	2,296	2,198	2,029
	Non-Hispanic Multiracial	Mean	353.51	334.74	310.85	284.46	319.29	323.05	340.46	321.76
		N	1,593	1,582	1,662	1,561	1,501	1,580	1,518	1,427
	Non-Hispanic White	Mean	352.37	332.72	312.45	284.80	318.87	322.73	338.70	321.40
		N	34,193	33,784	36,068	33,997	32,375	33,779	32,233	30,608
	Unknown	Mean	328.60	320.63	295.97	265.68	297.38	308.25	323.02	304.75
		N	26,273	26,412	27,951	26,092	24,650	26,403	25,026	23,557

Cluster	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
4–5	Non-Hispanic Asian	Mean	424.44	367.40	349.18	322.09	373.66	358.38	384.60	362.75
		N	30,653	30,132	30,683	30,150	28,774	28,942	28,841	26,277
	Non-Hispanic Pacific Islander	Mean	404.40	350.25	336.24	303.29	354.38	343.35	366.57	346.69
		N	2,603	2,528	2,626	2,581	2,420	2,419	2,399	2,149
	Non-Hispanic Black	Mean	414.71	354.78	335.36	321.58	368.55	345.21	372.84	352.18
		N	16,874	16,545	16,997	16,619	15,694	15,820	15,678	14,051
	Hispanic (Of Any Race)	Mean	412.84	354.88	336.84	312.08	362.82	345.93	372.32	350.85
		N	209,415	205,462	210,847	206,213	196,115	197,705	196,038	178,108
	Non-Hispanic American Indian	Mean	411.74	351.46	332.63	305.88	359.65	342.07	369.92	347.78
		N	2,122	2,115	2,184	2,111	1,962	2,022	1,974	1,765
	Non-Hispanic Multiracial	Mean	423.89	364.69	345.03	324.15	374.39	355.04	382.43	360.51
		N	1,153	1,106	1,145	1,111	1,072	1,065	1,069	972
	Non-Hispanic White	Mean	423.50	363.23	343.83	323.96	374.12	353.61	381.36	359.58
		N	25,786	24,867	25,539	25,466	24,114	23,627	23,697	21,290
	Unknown	Mean	403.97	350.03	329.60	302.02	353.14	339.57	366.22	343.38
		N	27,109	26,792	28,304	26,540	24,986	26,248	25,332	23,077
6–8	Non-Hispanic Asian	Mean	413.06	366.18	335.15	329.31	371.57	350.78	380.55	357.01
		N	25,394	24,815	25,314	25,273	23,770	23,490	23,554	21,216
	Non-Hispanic Pacific Islander	Mean	392.36	346.08	323.21	310.71	351.82	334.86	360.47	340.12
		N	2,679	2,642	2,699	2,687	2,459	2,456	2,448	2,125
	Non-Hispanic Black	Mean	402.15	352.01	320.47	325.67	364.32	336.44	367.63	345.00
		N	17,714	17,399	18,128	17,907	16,361	16,476	16,088	14,300
	Hispanic (Of Any Race)	Mean	397.58	348.43	323.13	312.44	355.38	335.83	363.38	341.58
		N	201,405	200,261	205,439	202,317	188,773	192,254	188,897	171,871
	Non-Hispanic American Indian	Mean	402.34	348.85	325.19	312.18	357.89	337.12	365.40	343.67
		N	2,654	2,716	2,751	2,669	2,431	2,588	2,501	2,200
	Non-Hispanic Multiracial	Mean	409.98	357.31	327.03	326.16	368.61	342.62	373.31	350.62
		N	892	896	911	904	842	857	839	764
	Non-Hispanic White	Mean	409.84	358.68	330.04	327.47	369.24	344.52	374.33	351.99
		N	23,030	22,537	23,157	23,034	21,374	21,321	21,176	18,894
	Unknown	Mean	393.80	345.60	316.52	306.81	350.72	331.05	360.32	336.87
		N	32,851	32,825	33,685	32,788	30,542	31,504	30,776	27,855
9–12	Non-Hispanic Asian	Mean	409.02	395.33	363.63	333.75	371.54	379.73	399.61	377.06
		N	28,513	27,714	28,822	28,044	26,605	26,759	26,531	24,196
	Non-Hispanic Pacific Islander	Mean	392.37	377.86	354.40	308.40	350.44	366.39	382.49	361.42
		N	2,381	2,331	2,418	2,374	2,225	2,234	2,200	1,991
	Non-Hispanic Black	Mean	396.06	381.29	348.51	323.50	359.99	365.09	386.07	363.49
		N	22,871	22,190	23,534	22,933	21,323	21,410	20,850	18,997
	Hispanic (Of Any Race)	Mean	391.79	376.19	348.15	306.93	349.64	362.38	381.09	358.45
		N	195,018	192,863	200,140	195,097	183,346	186,609	182,400	167,592
	Non-Hispanic American Indian	Mean	401.86	382.23	353.12	312.48	357.30	367.98	388.22	364.61
		N	2,186	2,166	2,256	2,189	2,061	2,114	2,050	1,895
	Non-Hispanic Multiracial	Mean	408.69	389.88	358.10	328.66	369.10	374.11	395.67	372.76
		N	718	691	737	714	681	675	661	615
	Non-Hispanic White	Mean	406.25	387.53	353.36	324.14	365.53	370.72	393.60	369.20
		N	23,355	22,700	23,884	23,284	21,867	21,938	21,529	19,712
	Unknown	Mean	389.80	373.62	342.09	300.77	345.65	357.87	378.76	354.06
		N	34,503	34,120	36,150	34,917	32,164	33,071	31,856	29,162

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 ELs.

Table 1.2.2.1

Mean Scale Scores by Grade, S501 Online

Grade		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Mean	320.27	287.07	255.15	255.02	287.90	271.26	297.08	276.11
	N	176,572	179,739	186,850	174,883	165,935	179,697	170,589	160,535
2	Mean	325.43	318.68	293.78	267.81	296.88	306.30	320.69	303.27
	N	183,889	184,150	194,128	182,404	173,381	184,069	175,334	165,651
3	Mean	350.51	333.06	314.00	282.99	316.95	323.70	338.36	321.45
	N	182,714	182,462	192,009	181,680	173,473	182,392	174,482	165,992
4	Mean	410.91	354.68	334.40	313.30	362.52	344.59	371.62	349.87
	N	174,730	171,235	175,857	171,751	163,102	164,493	163,186	147,625
5	Mean	418.14	358.34	342.14	313.98	366.36	350.28	376.33	354.89
	N	140,985	138,312	142,468	139,040	132,035	133,355	131,842	120,064
6	Mean	396.10	345.28	320.05	314.34	355.60	332.75	360.76	339.56
	N	114,021	113,310	116,458	114,305	106,490	108,666	106,592	96,583
7	Mean	399.19	350.92	324.15	314.56	357.28	337.60	365.66	343.49
	N	102,671	101,914	104,665	102,963	95,689	97,506	95,881	86,658
8	Mean	404.76	356.83	328.13	316.83	361.17	342.50	371.47	347.90
	N	89,927	88,867	90,961	90,311	84,373	84,774	83,806	75,984
9	Mean	386.04	368.81	339.98	301.08	343.87	354.48	374.24	351.18
	N	102,240	101,358	105,520	102,993	96,040	97,938	95,312	87,417
10	Mean	394.84	379.28	349.85	312.22	353.79	364.79	384.21	361.40
	N	81,296	79,762	83,442	81,135	76,119	77,134	75,461	69,142
11	Mean	400.77	386.23	355.98	319.14	360.23	371.41	390.84	367.96
	N	67,599	66,214	69,280	67,164	63,123	64,104	62,705	57,295
12	Mean	402.52	387.97	357.72	318.96	360.97	373.06	392.52	369.22
	N	58,410	57,441	59,699	58,260	54,990	55,634	54,599	50,306

Table 1.2.2.2

Mean Scale Scores by Grade by Gender, S501 Online

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	F	Mean	323.41	289.28	260.08	261.20	292.54	274.81	299.53	279.93
		N	80,990	82,091	85,531	80,410	76,474	82,077	78,079	73,856
	M	Mean	317.58	285.34	251.08	249.95	284.03	268.37	295.11	272.98
		N	90,826	92,738	96,259	89,729	84,983	92,710	87,877	82,312
	Missing	Mean	318.10	282.85	249.24	246.20	282.19	266.28	293.42	270.76
		N	4,756	4,910	5,060	4,744	4,478	4,910	4,633	4,367
2	F	Mean	327.71	320.37	299.16	273.71	300.90	309.80	322.50	306.81
		N	84,233	83,836	88,747	83,679	79,698	83,801	79,979	75,834
	M	Mean	323.69	317.41	289.30	262.88	293.61	303.47	319.32	300.38
		N	94,693	95,268	100,110	93,711	88,951	95,229	90,593	85,277
	Missing	Mean	319.69	314.81	288.08	261.30	290.74	301.69	316.39	298.28
		N	4,963	5,046	5,271	5,014	4,732	5,039	4,762	4,540
3	F	Mean	351.65	335.30	319.62	288.77	320.39	327.65	340.22	325.18
		N	83,036	82,516	87,083	82,610	79,010	82,489	79,077	75,421
	M	Mean	349.77	331.36	309.31	278.26	314.22	320.50	336.98	318.42
		N	94,692	94,913	99,676	94,045	89,684	94,872	90,602	85,961
	Missing	Mean	345.62	328.42	310.09	276.63	311.35	319.44	333.62	316.82
		N	4,986	5,033	5,250	5,025	4,779	5,031	4,803	4,610
4	F	Mean	410.39	356.28	339.67	317.16	364.12	348.03	372.57	352.68
		N	78,586	76,464	78,799	77,376	73,631	73,438	73,067	66,333
	M	Mean	411.49	353.46	330.17	310.25	361.35	341.87	370.95	347.66
		N	92,057	90,654	92,876	90,263	85,624	87,110	86,250	77,780
	Missing	Mean	407.56	351.98	329.23	307.43	357.82	340.78	368.87	345.62
		N	4,087	4,117	4,182	4,112	3,847	3,945	3,869	3,512
5	F	Mean	417.50	360.75	347.82	316.91	367.46	354.36	377.81	358.00
		N	62,042	60,574	62,424	61,212	58,265	58,365	57,896	52,896
	M	Mean	418.68	356.40	337.57	311.71	365.54	347.01	375.14	352.39
		N	75,039	73,879	76,075	73,969	70,117	71,263	70,276	63,833
	Missing	Mean	418.09	357.38	340.17	310.87	364.73	348.77	375.62	353.30
		N	3,904	3,859	3,969	3,859	3,653	3,727	3,670	3,335

Grade	Gender		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
6	F	Mean	394.79	348.00	324.88	316.17	355.75	336.55	362.24	342.18
		N	48,940	48,312	49,662	48,878	45,814	46,371	45,753	41,603
	M	Mean	397.11	343.17	316.14	312.87	355.45	329.69	359.59	337.41
		N	61,770	61,682	63,404	62,065	57,560	59,106	57,733	52,134
	Missing	Mean	396.53	345.01	322.50	314.92	356.34	334.00	360.64	340.51
		N	3,311	3,316	3,392	3,362	3,116	3,189	3,106	2,846
7	F	Mean	398.11	353.91	328.66	316.83	357.77	341.35	367.38	346.14
		N	43,465	42,916	43,973	43,330	40,465	41,036	40,601	36,691
	M	Mean	400.03	348.71	320.67	312.88	356.93	334.73	364.37	341.44
		N	56,281	56,104	57,686	56,685	52,498	53,689	52,565	47,502
	Missing	Mean	399.11	349.62	325.13	313.43	356.93	337.66	364.70	343.52
		N	2,925	2,894	3,006	2,948	2,726	2,781	2,715	2,465
8	F	Mean	403.76	359.90	332.31	319.77	362.07	346.11	373.21	350.55
		N	38,356	37,541	38,308	38,120	35,906	35,756	35,691	32,300
	M	Mean	405.53	354.58	325.02	314.80	360.55	339.80	370.17	345.89
		N	49,142	48,893	50,117	49,682	46,160	46,661	45,871	41,593
	Missing	Mean	404.90	354.97	326.67	312.34	359.45	340.94	370.50	346.82
		N	2,429	2,433	2,536	2,509	2,307	2,357	2,244	2,091
9	F	Mean	386.60	372.64	345.91	304.49	345.83	359.38	377.02	355.12
		N	42,270	41,579	43,306	42,352	39,748	40,178	39,320	36,092
	M	Mean	385.51	366.07	335.75	298.93	342.50	350.99	372.18	348.34
		N	57,130	56,948	59,202	57,707	53,620	55,019	53,378	48,931
	Missing	Mean	388.46	367.74	337.92	294.31	342.34	352.85	374.36	349.95
		N	2,840	2,831	3,012	2,934	2,672	2,741	2,614	2,394
10	F	Mean	394.19	382.09	354.45	314.63	354.58	368.53	385.93	364.17
		N	34,199	33,276	34,814	33,921	32,019	32,128	31,671	28,976
	M	Mean	395.06	377.07	346.38	310.58	353.14	361.92	382.75	359.25
		N	44,703	44,110	46,092	44,764	41,849	42,687	41,567	38,103
	Missing	Mean	400.12	381.00	349.91	308.81	354.75	365.56	387.04	362.48
		N	2,394	2,376	2,536	2,450	2,251	2,319	2,223	2,063
11	F	Mean	400.01	388.67	360.04	321.31	360.89	374.68	392.29	370.47
		N	29,269	28,488	29,784	28,898	27,295	27,552	27,100	24,715
	M	Mean	401.27	384.31	352.99	317.79	359.81	368.91	389.64	366.06
		N	36,035	35,473	37,126	35,978	33,681	34,352	33,470	30,607
	Missing	Mean	402.52	385.70	351.95	313.05	358.35	369.37	391.19	366.15
		N	2,295	2,253	2,370	2,288	2,147	2,200	2,135	1,973
12	F	Mean	402.17	390.82	362.06	321.77	362.09	376.66	394.36	371.97
		N	25,675	25,074	26,078	25,495	24,135	24,232	23,918	21,998
	M	Mean	402.60	385.52	354.11	316.87	360.02	370.02	390.85	366.87
		N	31,058	30,689	31,885	31,064	29,256	29,762	29,094	26,825
	Missing	Mean	406.27	390.22	358.79	315.11	361.35	374.95	395.31	370.83
		N	1,677	1,678	1,736	1,701	1,599	1,640	1,587	1,483

Table 1.2.2.3

Mean Scale Scores by Grade by Ethnicity, S501 Online

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
1	Non-Hispanic Asian	Mean	332.64	303.85	271.60	263.85	298.47	287.94	312.60	291.04
		N	24,293	24,495	25,567	24,002	22,924	24,491	23,398	22,112
	Non-Hispanic Pacific Islander	Mean	304.45	281.21	254.89	247.47	275.55	268.08	287.94	269.73
		N	1,531	1,571	1,635	1,534	1,439	1,570	1,483	1,398
	Non-Hispanic Black	Mean	321.41	290.20	258.01	267.61	294.81	274.31	299.71	280.53
		N	9,520	9,727	10,173	9,462	8,893	9,724	9,147	8,560
	Hispanic (Of Any Race)	Mean	316.70	282.63	250.64	251.50	284.33	266.79	292.89	271.87
		N	110,275	112,556	116,637	109,355	103,764	112,530	106,788	100,602
	Non-Hispanic American Indian	Mean	320.34	280.81	250.12	250.65	285.91	265.64	292.53	271.63
		N	1,175	1,205	1,238	1,141	1,088	1,204	1,150	1,065
	Non-Hispanic Multiracial	Mean	335.62	299.11	266.53	266.48	301.64	283.10	310.07	288.51
		N	782	780	824	772	734	779	742	696
	Non-Hispanic White	Mean	331.18	293.59	263.81	264.89	298.36	278.81	304.94	284.54
		N	16,805	17,001	17,837	16,668	15,771	16,997	16,118	15,153
	Unknown	Mean	312.96	283.41	248.91	246.39	280.10	266.26	292.26	270.31
		N	12,191	12,404	12,939	11,949	11,322	12,402	11,763	10,949
2	Non-Hispanic Asian	Mean	343.03	330.72	309.69	277.10	310.28	320.42	334.42	317.11
		N	24,054	24,089	25,337	23,917	22,815	24,084	23,018	21,892
	Non-Hispanic Pacific Islander	Mean	308.06	313.81	292.36	253.94	281.41	303.24	312.32	296.54
		N	1,548	1,577	1,671	1,543	1,441	1,577	1,480	1,382
	Non-Hispanic Black	Mean	327.35	319.23	295.56	277.52	302.79	307.52	321.73	305.98
		N	10,074	10,206	10,828	10,081	9,428	10,204	9,560	8,971
	Hispanic (Of Any Race)	Mean	320.34	315.66	289.67	264.82	292.81	302.73	317.03	299.52
		N	115,542	115,724	121,688	114,434	108,951	115,675	110,358	104,248
	Non-Hispanic American Indian	Mean	324.21	314.82	290.90	260.64	292.40	302.93	317.38	299.16
		N	1,134	1,146	1,188	1,122	1,056	1,130	1,085	1,000
	Non-Hispanic Multiracial	Mean	345.35	328.70	303.43	277.20	312.02	316.31	333.85	315.11
		N	820	811	858	805	770	809	776	726
	Non-Hispanic White	Mean	341.12	325.14	303.72	277.49	309.62	314.47	329.92	312.79
		N	17,440	17,208	18,420	17,325	16,483	17,207	16,399	15,545
	Unknown	Mean	316.63	314.79	286.11	258.39	287.85	300.42	315.39	296.55
		N	13,277	13,389	14,138	13,177	12,437	13,383	12,658	11,887

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
3	Non-Hispanic Asian	Mean	366.03	346.55	326.55	291.40	328.88	336.82	352.48	334.24
		N	21,449	21,496	22,451	21,340	20,477	21,492	20,633	19,736
	Non-Hispanic Pacific Islander	Mean	334.94	324.77	314.10	272.27	303.82	319.59	327.91	314.56
		N	1,537	1,550	1,655	1,579	1,471	1,550	1,456	1,395
	Non-Hispanic Black	Mean	350.91	333.72	314.56	290.26	320.79	324.26	339.00	323.04
		N	10,020	9,968	10,599	9,981	9,477	9,967	9,493	9,000
	Hispanic (Of Any Race)	Mean	346.96	330.35	311.53	280.78	314.07	321.11	335.37	318.74
		N	118,017	117,910	123,809	117,299	112,132	117,854	112,843	107,398
	Non-Hispanic American Indian	Mean	347.02	323.38	307.19	272.93	310.38	315.45	330.78	313.93
		N	1,169	1,168	1,230	1,138	1,080	1,166	1,113	1,029
	Non-Hispanic Multiracial	Mean	362.16	341.09	318.75	292.18	326.95	330.13	347.37	328.66
		N	773	771	804	756	731	771	742	701
	Non-Hispanic White	Mean	364.08	340.59	321.57	292.40	328.47	331.31	347.80	330.28
		N	16,753	16,576	17,648	16,672	15,892	16,572	15,834	15,063
	Unknown	Mean	340.82	326.62	306.06	273.12	307.08	316.29	330.82	313.11
		N	12,996	13,023	13,813	12,915	12,213	13,020	12,368	11,670
4	Non-Hispanic Asian	Mean	422.67	366.07	346.22	321.65	372.66	356.21	383.21	361.07
		N	18,212	17,921	18,200	17,887	17,070	17,195	17,173	15,589
	Non-Hispanic Pacific Islander	Mean	399.10	347.03	331.27	302.48	351.29	338.97	362.75	342.66
		N	1,458	1,420	1,475	1,461	1,363	1,347	1,348	1,202
	Non-Hispanic Black	Mean	412.16	353.55	332.96	320.80	366.92	343.49	371.22	350.50
		N	9,507	9,292	9,566	9,356	8,833	8,873	8,794	7,868
	Hispanic (Of Any Race)	Mean	409.10	352.96	333.01	311.66	360.78	343.06	369.86	348.26
		N	114,741	112,495	115,393	112,810	107,274	108,119	107,295	97,229
	Non-Hispanic American Indian	Mean	407.95	349.01	325.79	304.22	357.10	337.58	367.19	344.45
		N	1,036	1,041	1,070	1,030	960	1,001	974	874
	Non-Hispanic Multiracial	Mean	422.75	364.40	344.22	325.90	374.72	354.48	381.95	360.18
		N	639	615	637	617	595	592	592	539
	Non-Hispanic White	Mean	420.57	361.84	341.04	323.67	372.54	351.61	379.52	357.74
		N	14,813	14,303	14,613	14,607	13,838	13,540	13,636	12,205
	Unknown	Mean	400.48	348.19	325.62	301.23	351.05	336.61	363.86	340.67
		N	14,324	14,148	14,903	13,983	13,169	13,826	13,374	12,119

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
5	Non-Hispanic Asian	Mean	427.02	369.36	353.48	322.74	375.12	361.56	386.65	365.21
		N	12,441	12,211	12,483	12,263	11,704	11,747	11,668	10,688
	Non-Hispanic Pacific Islander	Mean	411.15	354.38	342.60	304.36	358.36	348.87	371.48	351.81
		N	1,145	1,108	1,151	1,120	1,057	1,072	1,051	947
	Non-Hispanic Black	Mean	418.01	356.35	338.46	322.59	370.66	347.41	374.91	354.31
		N	7,367	7,253	7,431	7,263	6,861	6,947	6,884	6,183
	Hispanic (Of Any Race)	Mean	417.37	357.19	341.47	312.59	365.28	349.40	375.29	353.96
		N	94,674	92,967	95,454	93,403	88,841	89,586	88,743	80,879
	Non-Hispanic American Indian	Mean	415.36	353.84	339.20	307.46	362.08	346.47	372.58	351.05
		N	1,086	1,074	1,114	1,081	1,002	1,021	1,000	891
	Non-Hispanic Multiracial	Mean	425.30	365.05	346.05	321.98	373.97	355.74	383.03	360.91
		N	514	491	508	494	477	473	477	433
	Non-Hispanic White	Mean	427.46	365.11	347.56	324.35	376.26	356.30	383.86	362.06
		N	10,973	10,564	10,926	10,859	10,276	10,087	10,061	9,085
	Unknown	Mean	407.88	352.08	334.02	302.89	355.47	342.87	368.86	346.37
		N	12,785	12,644	13,401	12,557	11,817	12,422	11,958	10,958
6	Non-Hispanic Asian	Mean	405.90	357.58	328.51	323.37	364.95	343.21	372.32	349.66
		N	8,994	8,841	9,074	8,955	8,424	8,464	8,378	7,602
	Non-Hispanic Pacific Islander	Mean	391.79	344.14	322.78	309.23	350.98	333.54	358.83	339.07
		N	1,099	1,068	1,115	1,105	1,020	1,012	1,001	891
	Non-Hispanic Black	Mean	396.96	344.85	314.97	322.38	360.05	330.11	360.96	339.24
		N	6,014	6,009	6,254	6,092	5,522	5,704	5,507	4,878
	Hispanic (Of Any Race)	Mean	394.80	343.81	319.79	312.56	354.04	331.87	359.31	338.45
		N	76,248	75,858	77,937	76,556	71,433	72,924	71,560	65,133
	Non-Hispanic American Indian	Mean	396.29	343.47	319.72	310.74	354.10	331.81	360.01	338.82
		N	974	980	994	979	898	935	911	808
	Non-Hispanic Multiracial	Mean	407.16	352.57	321.92	322.76	365.97	337.34	369.17	346.23
		N	361	363	370	365	342	348	341	308
	Non-Hispanic White	Mean	403.69	352.06	325.59	324.46	364.55	339.00	367.74	346.61
		N	8,545	8,309	8,604	8,530	7,949	7,891	7,806	6,995
	Unknown	Mean	391.16	341.07	313.79	308.07	350.07	327.47	356.33	334.21
		N	11,786	11,882	12,110	11,723	10,902	11,388	11,088	9,968

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
7	Non-Hispanic Asian	Mean	413.43	367.08	335.77	330.10	372.14	351.55	381.34	357.79
		N	8,505	8,326	8,491	8,507	7,961	7,851	7,900	7,100
	Non-Hispanic Pacific Islander	Mean	389.90	345.02	321.74	312.42	351.42	333.67	359.10	339.48
		N	878	863	865	875	801	784	800	673
	Non-Hispanic Black	Mean	401.14	352.30	320.69	325.44	363.73	336.82	367.56	345.10
		N	5,918	5,811	6,081	6,035	5,463	5,498	5,342	4,751
	Hispanic (Of Any Race)	Mean	397.03	348.76	323.56	311.56	354.70	336.21	363.44	341.71
		N	67,572	67,243	68,969	67,700	63,116	64,570	63,398	57,548
	Non-Hispanic American Indian	Mean	401.38	347.29	324.31	310.85	356.80	335.81	363.94	342.23
		N	918	947	960	921	837	902	865	752
	Non-Hispanic Multiracial	Mean	409.58	357.99	328.59	325.37	367.41	343.83	373.91	351.18
		N	298	300	308	300	279	290	280	258
	Non-Hispanic White	Mean	410.51	359.84	331.04	328.29	370.03	345.65	375.35	353.15
		N	7,627	7,511	7,728	7,680	7,058	7,112	7,048	6,297
	Unknown	Mean	392.79	345.69	316.21	305.62	349.58	330.96	360.06	336.38
		N	10,955	10,913	11,263	10,945	10,174	10,499	10,248	9,279
8	Non-Hispanic Asian	Mean	420.81	375.14	342.26	335.26	378.50	358.86	389.15	364.73
		N	7,895	7,648	7,749	7,811	7,385	7,175	7,276	6,514
	Non-Hispanic Pacific Islander	Mean	396.34	350.29	325.64	310.90	353.68	338.28	364.70	342.56
		N	702	711	719	707	638	660	647	561
	Non-Hispanic Black	Mean	408.58	359.41	326.17	329.36	369.32	342.89	374.71	350.91
		N	5,782	5,579	5,793	5,780	5,376	5,274	5,239	4,671
	Hispanic (Of Any Race)	Mean	401.90	354.20	327.08	313.32	357.94	340.67	368.72	345.58
		N	57,585	57,160	58,533	58,061	54,224	54,760	53,939	49,190
	Non-Hispanic American Indian	Mean	411.24	357.41	333.07	315.61	364.10	345.30	373.93	351.47
		N	762	789	797	769	696	751	725	640
	Non-Hispanic Multiracial	Mean	414.85	363.81	333.08	332.33	374.23	349.41	379.01	356.71
		N	233	233	233	239	221	219	218	198
	Non-Hispanic White	Mean	416.77	365.58	334.52	330.30	374.22	350.14	381.31	357.40
		N	6,858	6,717	6,825	6,824	6,367	6,318	6,322	5,602
	Unknown	Mean	397.98	350.86	320.08	306.64	352.70	335.37	365.29	340.49
		N	10,110	10,030	10,312	10,120	9,466	9,617	9,440	8,608

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
9	Non-Hispanic Asian	Mean	406.53	390.49	359.55	330.52	368.69	375.34	395.61	373.20
		N	7,961	7,786	8,031	7,845	7,389	7,476	7,422	6,698
	Non-Hispanic Pacific Islander	Mean	385.18	369.79	348.67	302.26	343.78	359.52	374.61	354.46
		N	807	796	824	820	759	756	743	668
	Non-Hispanic Black	Mean	393.15	374.02	342.11	318.37	356.24	358.23	380.23	357.70
		N	6,707	6,605	6,968	6,760	6,249	6,386	6,173	5,632
	Hispanic (Of Any Race)	Mean	382.12	365.42	337.81	296.02	339.37	351.71	370.63	347.90
		N	67,482	67,146	69,651	68,049	63,596	64,929	63,158	58,107
	Non-Hispanic American Indian	Mean	394.16	373.79	347.54	304.21	349.79	360.98	379.81	357.51
		N	788	778	798	790	752	752	742	687
	Non-Hispanic Multiracial	Mean	407.41	383.21	350.92	320.38	364.41	366.57	391.09	366.38
		N	225	220	238	232	216	216	206	195
	Non-Hispanic White	Mean	400.53	380.48	348.48	318.84	359.94	364.72	386.94	363.19
		N	7,158	6,975	7,332	7,164	6,690	6,722	6,592	6,011
	Unknown	Mean	380.61	362.95	331.56	288.88	335.13	347.00	368.44	343.20
		N	11,112	11,052	11,678	11,333	10,389	10,701	10,276	9,419
10	Non-Hispanic Asian	Mean	409.13	395.55	363.59	334.50	371.98	379.81	399.79	377.25
		N	7,341	7,104	7,414	7,203	6,850	6,873	6,818	6,221
	Non-Hispanic Pacific Islander	Mean	392.38	376.00	353.27	307.89	349.87	364.60	380.80	359.41
		N	661	637	672	660	620	614	603	550
	Non-Hispanic Black	Mean	397.08	382.15	348.90	324.41	360.81	365.83	386.97	364.37
		N	5,736	5,480	5,841	5,690	5,330	5,272	5,177	4,704
	Hispanic (Of Any Race)	Mean	391.97	376.70	348.89	308.28	350.37	363.03	381.46	359.09
		N	51,837	51,129	53,168	51,773	48,657	49,454	48,392	44,416
	Non-Hispanic American Indian	Mean	404.86	383.34	354.46	315.73	359.80	368.97	390.12	365.80
		N	580	576	600	573	543	566	549	505
	Non-Hispanic Multiracial	Mean	401.99	387.40	356.89	328.87	365.67	372.73	391.85	370.96
		N	189	176	188	179	175	170	171	156
	Non-Hispanic White	Mean	407.00	388.55	353.82	325.52	366.67	371.49	394.71	370.27
		N	5,983	5,746	6,108	5,964	5,601	5,545	5,454	4,988
	Unknown	Mean	389.61	373.23	341.85	300.44	345.36	357.54	378.52	353.93
		N	8,969	8,914	9,451	9,093	8,343	8,640	8,297	7,602

Grade	Ethnicity		Listening	Reading	Writing	Speaking	Oral	Literacy	Compre-hension	Overall
11	Non-Hispanic Asian	Mean	410.09	397.60	365.19	334.74	372.52	381.72	401.52	378.74
		N	7,061	6,855	7,148	6,971	6,626	6,626	6,572	6,030
	Non-Hispanic Pacific Islander	Mean	399.91	387.46	361.22	317.71	359.40	374.62	391.75	370.10
		N	490	474	491	478	455	455	457	411
	Non-Hispanic Black	Mean	398.08	385.60	352.42	327.52	362.80	369.23	389.75	367.33
		N	5,196	5,010	5,358	5,217	4,829	4,834	4,687	4,243
	Hispanic (Of Any Race)	Mean	399.18	384.59	355.97	315.90	357.78	370.57	389.14	366.61
		N	40,924	40,251	41,868	40,649	38,302	38,979	38,159	34,936
	Non-Hispanic American Indian	Mean	406.95	390.86	357.84	320.18	364.33	375.12	396.48	372.21
		N	456	442	472	450	419	431	417	377
	Non-Hispanic Multiracial	Mean	414.02	397.58	365.22	341.45	378.35	381.63	401.84	380.53
		N	160	155	165	160	153	153	150	141
	Non-Hispanic White	Mean	410.22	391.88	357.59	329.56	370.45	375.06	397.69	373.73
		N	5,268	5,153	5,382	5,227	4,918	4,983	4,892	4,462
	Unknown	Mean	395.65	380.89	348.87	309.35	352.89	365.12	385.64	361.27
		N	8,044	7,874	8,396	8,012	7,421	7,643	7,371	6,695
12	Non-Hispanic Asian	Mean	410.88	398.79	367.16	335.90	373.57	383.05	402.39	379.83
		N	6,150	5,969	6,229	6,025	5,740	5,784	5,719	5,247
	Non-Hispanic Pacific Islander	Mean	397.35	385.07	359.35	310.63	353.82	372.61	389.13	367.45
		N	423	424	431	416	391	409	397	362
	Non-Hispanic Black	Mean	396.65	385.57	352.50	325.10	361.12	369.13	389.00	366.24
		N	5,232	5,095	5,367	5,266	4,915	4,918	4,813	4,418
	Hispanic (Of Any Race)	Mean	401.61	386.66	358.15	315.82	358.96	372.64	391.34	368.37
		N	34,775	34,337	35,453	34,626	32,791	33,247	32,691	30,133
	Non-Hispanic American Indian	Mean	407.38	387.98	356.82	315.69	361.19	372.45	393.33	368.95
		N	362	370	386	376	347	365	342	326
	Non-Hispanic Multiracial	Mean	413.57	394.96	363.33	327.52	370.53	379.32	400.69	376.24
		N	144	140	146	143	137	136	134	123
	Non-Hispanic White	Mean	409.38	391.88	355.40	324.42	367.00	373.80	397.50	371.71
		N	4,946	4,826	5,062	4,929	4,658	4,688	4,591	4,251
	Unknown	Mean	398.67	383.82	352.39	311.41	355.30	368.36	388.45	364.15
		N	6,378	6,280	6,625	6,479	6,011	6,087	5,912	5,446

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 Cluster 1, Listening was correlated to Speaking and Writing; Reading was correlated to Writing. In Cluster 2–3, Listening was mostly correlated to Writing, and Reading was also correlated to Writing. In Clusters 4–5, 6–8, and 9–12, the Listening and Reading domains were highly correlated. The Writing domain was also correlated to the Reading/Listening domain.

Table 1.2.3.1

Correlations Among Scale Scores: Grade 1, S501 Online

		Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1	0.419	0.512	0.519
	N	176,572	170,589	176,530	165,935
Reading	Pearson Correlation		1	0.484	0.321
	N		179,739	179,697	168,688
Writing	Pearson Correlation			1	0.428
	N			186,850	174,840
Speaking	Pearson Correlation				1
	N				174,883

Table 1.2.3.2

Correlations Among Scale Scores: Grades 2–3, S501 Online

		Listening	Reading	Writing	Speaking
Listening	Pearson	1	0.597	0.646	0.604
	N	366,603	349,816	366,449	346,854
Reading	Pearson Correlation		1	0.617	0.497
	N		366,612	366,461	346,838
Writing	Pearson Correlation			1	0.600
	N			386,137	363,940
Speaking	Pearson Correlation				1
	N				364,084

Table 1.2.3.3

Correlations Among Scale Scores: Grades 4–5, S501 Online

		Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1	0.664	0.646	0.606
	N	315,715	295,028	302,166	295,137
Reading	Pearson Correlation		1	0.635	0.525
	N		309,547	297,848	289,908
Writing	Pearson			1	0.603
	N			318,325	297,435
Speaking	Pearson Correlation				1
	N				310,791

Table 1.2.3.4

Correlations Among Scale Scores: Grades 6–8, S501 Online

		Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1	0.680	0.616	0.587
	N	306,619	286,279	291,667	286,552
Reading	Pearson Correlation		1	0.674	0.543
	N		304,091	290,946	284,756
Writing	Pearson Correlation			1	0.602
	N			312,084	291,923
Speaking	Pearson Correlation				1
	N				307,579

Table 1.2.3.5

Correlations Among Scale Scores: Grades 9–12, S501 Online

		Listening	Reading	Writing	Speaking
Listening	Pearson Correlation	1	0.719	0.530	0.584
	N	309,545	288,077	298,041	290,272
Reading	Pearson Correlation		1	0.604	0.610
	N		304,775	294,810	286,884
Writing	Pearson Correlation			1	0.608
	N			317,941	297,848
Speaking	Pearson Correlation				1
	N				309,552

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 in Proficiency Level (PL) 6, especially in Cluster 4–5. Cluster 1, 2–3, and 6–8 also had over 40% in PL 6. The Reading domain had 7% to 17% in PL 6. For the Writing domain, fewer than 1% of students were in PL 5 and PL 6 together; Cluster 4–5 showed 3% in both PL ranges. 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 1.3.1.1.1.1

Proficiency Level by Cluster (Count): Listening, S501 Online

Cluster	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	17,640	8,433	21,438	13,344	22,142	93,575	176,572
2–3	40,176	40,470	60,852	21,850	52,423	150,832	366,603
4–5	9,217	7,742	18,079	11,608	35,114	233,955	315,715
6–8	9,824	22,386	44,236	47,608	54,205	128,360	306,619
9–12	27,633	39,962	68,901	70,050	49,983	53,016	309,545

Table 1.3.1.1.1.2

Proficiency Level by Cluster (Percent): Listening, S501 Online

Cluster	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	10.0%	4.8%	12.1%	7.6%	12.5%	53.0%	100.0%
2–3	11.0%	11.0%	16.6%	6.0%	14.3%	41.1%	100.0%
4–5	2.9%	2.5%	5.7%	3.7%	11.1%	74.1%	100.0%
6–8	3.2%	7.3%	14.4%	15.5%	17.7%	41.9%	100.0%
9–12	8.9%	12.9%	22.3%	22.6%	16.1%	17.1	100.0%

1.3.1.1.2 By Grade

Table 1.3.1.1.2.1

Proficiency Level by Grade (Count): Listening, S501 Online

Grade	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	17,640	8,433	21,438	13,344	22,142	93,575	176,572
2	23,030	22,404	32,675	10,972	21,961	72,847	183,889
3	17,146	18,066	28,177	10,878	30,462	77,985	182,714
4	4,043	3,731	9,961	6,607	19,097	131,291	174,730
5	5,174	4,011	8,118	5,001	16,017	102,664	140,985
6	2,445	6,158	15,550	16,549	23,157	50,162	114,021
7	3,694	7,468	15,298	17,324	17,672	41,215	102,671
8	3,685	8,760	13,388	13,735	13,376	36,983	89,927
9	7,709	15,378	23,214	23,049	15,053	17,837	102,240
10	7,105	10,416	17,422	18,027	14,138	14,188	81,296
11	6,059	8,323	15,222	14,502	11,084	12,409	67,599
12	6,760	5,845	13,043	14,472	9,708	8,582	58,410

Table 1.3.1.1.2.2

Proficiency Level by Grade (Percent): Listening, S501 Online

Grade	Listening Proficiency Range						Total
	1	2	3	4	5	6	
1	10.0%	4.8%	12.1%	7.6%	12.5%	53.0%	100.0%
2	12.5%	12.2%	17.8%	6.0%	11.9%	39.6%	100.0%
3	9.4%	9.9%	15.4%	6.0%	16.7%	42.7%	100.0%
4	2.3%	2.1%	5.7%	3.8%	10.9%	75.1%	100.0%
5	3.7%	2.8%	5.8%	3.5%	11.4%	72.8%	100.0%
6	2.1%	5.4%	13.6%	14.5%	20.3%	44.0%	100.0%
7	3.6%	7.3%	14.9%	16.9%	17.2%	40.1%	100.0%
8	4.1%	9.7%	14.9%	15.3%	14.9%	41.1%	100.0%
9	7.5%	15.0%	22.7%	22.5%	14.7%	17.4%	100.0%
10	8.7%	12.8%	21.4%	22.2%	17.4%	17.5%	100.0%
11	9.0%	12.3%	22.5%	21.5%	16.4%	18.4%	100.0%
12	11.6%	10.0%	22.3%	24.8%	16.6%	14.7%	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, S501 Online

Cluster	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	40,892	51,970	37,854	17,380	15,046	16,597	179,739
2–3	46,661	91,497	71,350	46,976	61,306	48,822	366,612
4–5	33,361	60,728	50,979	43,059	66,669	54,751	309,547
6–8	92,132	74,508	69,492	15,947	29,910	22,102	304,091
9–12	69,001	81,417	50,446	18,906	46,146	38,859	304,775

Table 1.3.1.2.1.2

Proficiency Level by Cluster (Percent): Reading, S501 Online

Cluster	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	22.8%	28.9%	21.1%	9.7%	8.4%	9.2%	100.0%
2–3	12.7%	25.0%	19.5%	12.8%	16.7%	13.3%	100.0%
4–5	10.8%	19.6%	16.5%	13.9%	21.5%	17.7%	100.0%
6–8	30.3%	24.5%	22.9%	5.2%	9.8%	7.3%	100.0%
9–12	22.6%	26.7%	16.6%	6.2%	15.1%	12.8%	100.0%

1.3.1.2.2 By Grade

Table 1.3.1.2.2.1

Proficiency Level by Grade (Count): Reading, S501 Online

Grade	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	40,892	51,970	37,854	17,380	15,046	16,597	179,739
2	20,317	45,326	39,653	29,979	28,646	20,229	184,150
3	26,344	46,171	31,697	16,997	32,660	28,593	182,462
4	15,469	32,720	24,694	28,221	38,073	32,058	171,235
5	17,892	28,008	26,285	14,838	28,596	22,693	138,312
6	33,187	26,975	29,500	7,075	11,225	5,348	113,310
7	30,678	27,049	21,495	4,809	10,298	7,585	101,914
8	28,267	20,484	18,497	4,063	8,387	9,169	88,867
9	27,945	27,383	16,535	5,075	12,904	11,516	101,358
10	17,220	21,149	13,101	4,848	12,236	11,208	79,762
11	12,360	17,140	11,407	4,460	11,340	9,507	66,214
12	11,476	15,745	9,403	4,523	9,666	6,628	57,441

Table 1.3.1.2.2.2

Proficiency Level by Grade (Percent): Reading, S501 Online

Grade	Reading Proficiency Range						Total
	1	2	3	4	5	6	
1	22.8%	28.9%	21.1%	9.7%	8.4%	9.2%	100.0%
2	11.0%	24.6%	21.5%	16.3%	15.6%	11.0%	100.0%
3	14.4%	25.3%	17.4%	9.3%	17.9%	15.7%	100.0%
4	9.0%	19.1%	14.4%	16.5%	22.2%	18.7%	100.0%
5	12.9%	20.2%	19.0%	10.7%	20.7%	16.4%	100.0%
6	29.3%	23.8%	26.0%	6.2%	9.9%	4.7%	100.0%
7	30.1%	26.5%	21.1%	4.7%	10.1%	7.4%	100.0%
8	31.8%	23.1%	20.8%	4.6%	9.4%	10.3%	100.0%
9	27.6%	27.0%	16.3%	5.0%	12.7%	11.4%	100.0%
10	21.6%	26.5%	16.4%	6.1%	15.3%	14.1%	100.0%
11	18.7%	25.9%	17.2%	6.7%	17.1%	14.4%	100.0%
12	20.0%	27.4%	16.4%	7.9%	16.8%	11.5%	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, S501 Online

Cluster	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	41,942	101,286	42,387	1,203	32	0	186,850
2–3	30,944	66,716	245,539	42,631	298	9	386,137
4–5	16,031	17,351	153,803	121,281	8,550	1,309	318,325
6–8	27,237	58,376	182,481	43,661	312	17	312,084
9–12	35,388	52,075	174,147	55,104	1,214	13	317,941

Table 1.3.1.3.1.2

Proficiency Level by Cluster (Percent): Writing, S501 Online

Cluster	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	22.4%	54.2%	22.7%	0.6%	0.0%	0.0%	100.0%
2–3	8.0%	17.3%	63.6%	11.0%	0.1%	0.0%	100.0%
4–5	5.0%	5.5%	48.3%	38.1%	2.7%	0.4%	100.0%
6–8	8.7%	18.7%	58.5%	14.0%	0.1%	0.0%	100.0%
9–12	11.1%	16.4%	54.8%	17.3%	0.4%	0.0%	100.0%

1.3.1.3.2 By Grade

Table 1.3.1.3.2.1

Proficiency Level by Grade (Count): Writing, S501 Online

Grade	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	41,942	101,286	42,387	1,203	32	0	186,850
2	18,746	44,915	120,271	10,159	35	2	194,128
3	12,198	21,801	125,268	32,472	263	7	192,009
4	9,271	9,605	93,506	59,639	2,941	895	175,857
5	6,760	7,746	60,297	61,642	5,609	414	142,468
6	7,376	17,912	76,120	14,926	119	5	116,458
7	9,124	24,099	52,153	19,239	48	2	104,665
8	10,737	16,365	54,208	9,496	145	10	90,961
9	13,845	17,051	50,825	23,466	322	11	105,520
10	7,612	11,683	52,941	10,817	388	1	83,442
11	6,563	14,120	36,991	11,194	411	1	69,280
12	7,368	9,221	33,390	9,627	93	0	59,699

Table 1.3.1.3.2.2

Proficiency Level by Grade (Percent): Writing, S501 Online

Grade	Writing Proficiency Range						Total
	1	2	3	4	5	6	
1	22.4%	54.2%	22.7%	0.6%	0.0%	0.0%	100.0%
2	9.7%	23.1%	62.0%	5.2%	0.0%	0.0%	100.0%
3	6.4%	11.4%	65.2%	16.9%	0.1%	0.0%	100.0%
4	5.3%	5.5%	53.2%	33.9%	1.7%	0.5%	100.0%
5	4.7%	5.4%	42.3%	43.3%	3.9%	0.3%	100.0%
6	6.3%	15.4%	65.4%	12.8%	0.1%	0.0%	100.0%
7	8.7%	23.0%	49.8%	18.4%	0.0%	0.0%	100.0%
8	11.8%	18.0%	59.6%	10.4%	0.2%	0.0%	100.0%
9	13.1%	16.2%	48.2%	22.2%	0.3%	0.0%	100.0%
10	9.1%	14.0%	63.4%	13.0%	0.5%	0.0%	100.0%
11	9.5%	20.4%	53.4%	16.2%	0.6%	0.0%	100.0%
12	12.3%	15.4%	55.9%	16.1%	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, S501 Online

Cluster	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	28,880	59,964	59,060	25,048	1,843	88	174,883
2–3	50,682	115,293	161,390	34,608	1,735	376	364,084
4–5	28,621	70,532	134,138	73,023	4,120	357	310,791
6–8	56,778	79,157	134,186	36,318	1,086	54	307,579
9–12	101,954	79,931	117,005	10,342	220	100	309,552

Table 1.3.1.4.1.2

Proficiency Level by Cluster (Percent): Speaking, S501 Online

Cluster	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	16.5%	34.3%	33.8%	14.3%	1.1%	0.1%	100.0%
2–3	13.9%	31.7%	44.3%	9.5%	0.5%	0.1%	100.0%
4–5	9.2%	22.7%	43.2%	23.5%	1.3%	0.1%	100.0%
6–8	18.5%	25.7%	43.6%	11.8%	0.4%	0.0%	100.0%
9–12	32.9%	25.8%	37.8%	3.3%	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, S501 Online

Grade	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	28,880	59,964	59,060	25,048	1,843	88	174,883
2	25,938	68,138	71,128	16,164	939	97	182,404
3	24,744	47,155	90,262	18,444	796	279	181,680
4	13,309	42,255	74,838	39,131	1,944	274	171,751
5	15,312	28,277	59,300	33,892	2,176	83	139,040
6	15,831	33,022	50,117	14,888	436	11	114,305
7	18,512	24,594	49,457	10,151	233	16	102,963
8	22,435	21,541	34,612	11,279	417	27	90,311
9	40,901	21,801	36,278	3,947	47	19	102,993
10	25,297	17,726	35,488	2,542	54	28	81,135
11	19,703	20,801	23,887	2,686	56	31	67,164
12	16,053	19,603	21,352	1,167	63	22	58,260

Table 1.3.1.4.2.2

Proficiency Level by Grade (Percent): Speaking, S501 Online

Grade	Speaking Proficiency Range						Total
	1	2	3	4	5	6	
1	16.5%	34.3%	33.8%	14.3%	1.1%	0.1%	100.0%
2	14.2%	37.4%	39.0%	8.9%	0.5%	0.1%	100.0%
3	13.6%	26.0%	49.7%	10.2%	0.4%	0.2%	100.0%
4	7.7%	24.6%	43.6%	22.8%	1.1%	0.2%	100.0%
5	11.0%	20.3%	42.6%	24.4%	1.6%	0.1%	100.0%
6	13.8%	28.9%	43.8%	13.0%	0.4%	0.0%	100.0%
7	18.0%	23.9%	48.0%	9.9%	0.2%	0.0%	100.0%
8	24.8%	23.9%	38.3%	12.5%	0.5%	0.0%	100.0%
9	39.7%	21.2%	35.2%	3.8%	0.0%	0.0%	100.0%
10	31.2%	21.8%	43.7%	3.1%	0.1%	0.0%	100.0%
11	29.3%	31.0%	35.6%	4.0%	0.1%	0.0%	100.0%
12	27.6%	33.6%	36.6%	2.0%	0.1%	0.0%	100.0%

1.3.2 Composites

The observed order of performance of composite domains by percentages in PL5 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, S501 Online

Cluster	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	17,652	21,522	53,619	45,374	24,283	3,485	165,935
2–3	35,202	58,850	103,082	111,503	36,003	2,214	346,854
4–5	13,462	15,802	48,377	110,960	81,397	25,139	295,137
6–8	22,574	36,854	97,483	101,373	24,990	3,278	286,552
9–12	54,773	58,997	114,235	55,122	6,406	739	290,272

Table 1.3.2.1.1.2

Proficiency Level by Cluster (Percent): Oral, S501 Online

Cluster	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	10.6%	13.0%	32.3%	27.3%	14.6%	2.1%	100.0%
2–3	10.1%	17.0%	29.7%	32.1%	10.4%	0.6%	100.0%
4–5	4.6%	5.4%	16.4%	37.6%	27.6%	8.5%	100.0%
6–8	7.9%	12.9%	34.0%	35.4%	8.7%	1.1%	100.0%
9–12	18.9%	20.3%	39.4%	19.0%	2.2%	0.3%	100.0%

1.3.2.1.2 By Grade

Table 1.3.2.1.2.1

Proficiency Level by Grade (Count): Oral, S501 Online

Grade	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	17,652	21,522	53,619	45,374	24,283	3,485	165,935
2	18,767	32,934	54,968	48,483	17,043	1,186	173,381
3	16,435	25,916	48,114	63,020	18,960	1,028	173,473
4	6,230	8,326	25,671	60,868	46,332	15,675	163,102
5	7,232	7,476	22,706	50,092	35,065	9,464	132,035
6	5,767	11,448	37,171	41,256	9,778	1,070	106,490
7	7,973	13,041	33,298	31,860	8,397	1,120	95,689
8	8,834	12,365	27,014	28,257	6,815	1,088	84,373
9	20,106	21,286	33,411	18,761	2,231	245	96,040
10	13,763	14,993	30,207	15,150	1,796	210	76,119
11	10,810	12,124	26,651	11,903	1,459	176	63,123
12	10,094	10,594	23,966	9,308	920	108	54,990

Table 1.3.2.1.2.2

Proficiency Level by Grade (Percent): Oral, S501 Online

Grade	Oral Language Proficiency Range						Total
	1	2	3	4	5	6	
1	10.6%	13.0%	32.3%	27.3%	14.6%	2.1%	100.0%
2	10.8%	19.0%	31.7%	28.0%	9.8%	0.7%	100.0%
3	9.5%	14.9%	27.7%	36.3%	10.9%	0.6%	100.0%
4	3.8%	5.1%	15.7%	37.3%	28.4%	9.6%	100.0%
5	5.5%	5.7%	17.2%	37.9%	26.6%	7.2%	100.0%
6	5.4%	10.8%	34.9%	38.7%	9.2%	1.0%	100.0%
7	8.3%	13.6%	34.8%	33.3%	8.8%	1.2%	100.0%
8	10.5%	14.7%	32.0%	33.5%	8.1%	1.3%	100.0%
9	20.9%	22.2%	34.8%	19.5%	2.3%	0.3%	100.0%
10	18.1%	19.7%	39.7%	19.9%	2.4%	0.3%	100.0%
11	17.1%	19.2%	42.2%	18.9%	2.3%	0.3%	100.0%
12	18.4%	19.3%	43.6%	16.9%	1.7%	0.2%	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, S501 Online

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	39,835	76,171	52,189	9,786	1,584	132	179,697
2–3	32,816	66,426	173,812	84,173	8,794	440	366,461
4–5	21,017	26,201	111,465	108,173	26,289	4,703	297,848
6–8	48,087	69,273	128,709	41,284	3,394	199	290,946
9–12	40,362	65,980	121,572	56,460	10,120	316	294,810

Table 1.3.2.2.1.2

Proficiency Level by Cluster (Percent): Literacy, S501 Online

Cluster	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	22.2%	42.4%	29.0%	5.4%	0.9%	0.1%	100.0%
2–3	9.0%	18.1%	47.4%	23.0%	2.4%	0.1%	100.0%
4–5	7.1%	8.8%	37.4%	36.3%	8.8%	1.6%	100.0%
6–8	16.5%	23.8%	44.2%	14.2%	1.2%	0.1%	100.0%
9–12	13.7%	22.4%	41.2%	19.2%	3.4%	0.1%	100.0%

1.3.2.2.2 By Grade

Table 1.3.2.2.2.1

Proficiency Level by Grade (Count): Literacy, S501 Online

Grade	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	39,835	76,171	52,189	9,786	1,584	132	179,697
2	17,551	39,394	89,854	34,170	2,910	190	184,069
3	15,265	27,032	83,958	50,003	5,884	250	182,392
4	10,786	13,427	63,013	60,551	13,972	2,744	164,493
5	10,231	12,774	48,452	47,622	12,317	1,959	133,355
6	15,411	24,911	53,711	13,790	779	64	108,666
7	15,882	24,335	42,562	13,379	1,294	54	97,506
8	16,794	20,027	32,436	14,115	1,321	81	84,774
9	16,961	20,708	38,439	18,024	3,629	177	97,938
10	9,579	16,399	31,941	16,098	3,021	96	77,134
11	7,024	14,392	27,290	13,124	2,243	31	64,104
12	6,798	14,481	23,902	9,214	1,227	12	55,634

Table 1.3.2.2.2.2

Proficiency Level by Grade (Percent): Literacy, S501 Online

Grade	Literacy Proficiency Range						Total
	1	2	3	4	5	6	
1	22.2%	42.4%	29.0%	5.4%	0.9%	0.1%	100.0%
2	9.5%	21.4%	48.8%	18.6%	1.6%	0.1%	100.0%
3	8.4%	14.8%	46.0%	27.4%	3.2%	0.1%	100.0%
4	6.6%	8.2%	38.3%	36.8%	8.5%	1.7%	100.0%
5	7.7%	9.6%	36.3%	35.7%	9.2%	1.5%	100.0%
6	14.2%	22.9%	49.4%	12.7%	0.7%	0.1%	100.0%
7	16.3%	25.0%	43.7%	13.7%	1.3%	0.1%	100.0%
8	19.8%	23.6%	38.3%	16.7%	1.6%	0.1%	100.0%
9	17.3%	21.1%	39.2%	18.4%	3.7%	0.2%	100.0%
10	12.4%	21.3%	41.4%	20.9%	3.9%	0.1%	100.0%
11	11.0%	22.5%	42.6%	20.5%	3.5%	0.0%	100.0%
12	12.2%	26.0%	43.0%	16.6%	2.2%	0.0%	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, S501 Online

Cluster	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	16,876	31,849	44,059	23,043	29,925	24,837	170,589
2–3	31,492	63,676	74,860	45,274	69,305	65,209	349,816
4–5	12,872	25,713	39,681	35,496	70,372	110,894	295,028
6–8	39,933	57,369	65,675	45,598	43,794	33,910	286,279
9–12	41,829	70,368	60,334	33,860	44,014	37,672	288,077

Table 1.3.2.3.1.2

Proficiency Level by Cluster (Percent): Comprehension, S501 Online

Cluster	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	9.9%	18.7%	25.8%	13.5%	17.5%	14.6%	100.0%
2–3	9.0%	18.2%	21.4%	12.9%	19.8%	18.6%	100.0%
4–5	4.4%	8.7%	13.4%	12.0%	23.9%	37.6%	100.0%
6–8	13.9%	20.0%	22.9%	15.9%	15.3%	11.8%	100.0%
9–12	14.5%	24.4%	20.9%	11.8%	15.3%	13.1%	100.0%

1.3.2.3.2 By Grade

Table 1.3.2.3.2.1

Proficiency Level by Grade (Count): Comprehension, S501 Online

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	16,876	31,849	44,059	23,043	29,925	24,837	170,589
2	13,645	34,561	39,888	24,952	33,433	28,855	175,334
3	17,847	29,115	34,972	20,322	35,872	36,354	174,482
4	4,983	13,615	21,759	18,710	40,032	64,087	163,186
5	7,889	12,098	17,922	16,786	30,340	46,807	131,842
6	11,701	21,703	26,028	18,726	18,164	10,270	106,592
7	13,800	19,484	22,508	15,056	13,590	11,443	95,881
8	14,432	16,182	17,139	11,816	12,040	12,197	83,806
9	15,373	25,333	19,647	10,358	12,956	11,645	95,312
10	10,617	17,652	15,998	8,979	11,453	10,762	75,461
11	8,206	14,328	13,114	7,371	10,700	8,986	62,705
12	7,633	13,055	11,575	7,152	8,905	6,279	54,599

Table 1.3.2.3.2.2

Proficiency Level by Grade (Percent): Comprehension, S501 Online

Grade	Comprehension Proficiency Range						Total
	1	2	3	4	5	6	
1	9.9%	18.7%	25.8%	13.5%	17.5%	14.6%	100.0%
2	7.8%	19.7%	22.7%	14.2%	19.1%	16.5%	100.0%
3	10.2%	16.7%	20.0%	11.6%	20.6%	20.8%	100.0%
4	3.1%	8.3%	13.3%	11.5%	24.5%	39.3%	100.0%
5	6.0%	9.2%	13.6%	12.7%	23.0%	35.5%	100.0%
6	11.0%	20.4%	24.4%	17.6%	17.0%	9.6%	100.0%
7	14.4%	20.3%	23.5%	15.7%	14.2%	11.9%	100.0%
8	17.2%	19.3%	20.5%	14.1%	14.4%	14.6%	100.0%
9	16.1%	26.6%	20.6%	10.9%	13.6%	12.2%	100.0%
10	14.1%	23.4%	21.2%	11.9%	15.2%	14.3%	100.0%
11	13.1%	22.8%	20.9%	11.8%	17.1%	14.3%	100.0%
12	14.0%	23.9%	21.2%	13.1%	16.3%	11.5%	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, S501 Online

Cluster	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	20,548	50,453	72,193	14,561	2,635	145	160,535
2–3	29,325	57,762	144,994	90,013	9,355	194	331,643
4–5	15,198	19,352	77,498	116,333	34,596	4,712	267,689
6–8	29,459	53,540	116,652	54,886	4,458	230	259,225
9–12	39,954	55,707	113,584	49,440	5,310	165	264,160

Table 1.3.2.4.1.2

Proficiency Level by Cluster (Percent): Overall, S501 Online

Cluster	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	12.8%	31.4%	45.0%	9.1%	1.6%	0.1%	100.0%
2–3	8.8%	17.4%	43.7%	27.1%	2.8%	0.1%	100.0%
4–5	5.7%	7.2%	29.0%	43.5%	12.9%	1.8%	100.0%
6–8	11.4%	20.7%	45.0%	21.2%	1.7%	0.1%	100.0%
9–12	15.1%	21.1%	43.0%	18.7%	2.0%	0.1%	100.0%

1.3.2.4.2 By Grade

Table 1.3.2.4.2.1

Proficiency Level by Grade (Count): Overall, S501 Online

Grade	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	20,548	50,453	72,193	14,561	2,635	145	160,535
2	15,215	34,694	75,027	37,016	3,599	100	165,651
3	14,110	23,068	69,967	52,997	5,756	94	165,992
4	7,483	9,771	43,058	64,888	19,522	2,903	147,625
5	7,715	9,581	34,440	51,445	15,074	1,809	120,064
6	8,233	18,758	48,387	19,963	1,166	76	96,583
7	10,163	18,420	38,374	18,122	1,514	65	86,658
8	11,063	16,362	29,891	16,801	1,778	89	75,984
9	16,153	17,605	35,542	16,073	1,954	90	87,417
10	9,660	13,962	29,923	14,023	1,526	48	69,142
11	7,209	11,847	25,773	11,245	1,199	22	57,295
12	6,932	12,293	22,346	8,099	631	5	50,306

Table 1.3.2.4.2.2

Proficiency Level by Grade (Percent): Overall, S501 Online

Grade	Overall Proficiency Range						Total
	1	2	3	4	5	6	
1	12.8%	31.4%	45.0%	9.1%	1.6%	0.1%	100.0%
2	9.2%	20.9%	45.3%	22.3%	2.2%	0.1%	100.0%
3	8.5%	13.9%	42.2%	31.9%	3.5%	0.1%	100.0%
4	5.1%	6.6%	29.2%	44.0%	13.2%	2.0%	100.0%
5	6.4%	8.0%	28.7%	42.8%	12.6%	1.5%	100.0%
6	8.5%	19.4%	50.1%	20.7%	1.2%	0.1%	100.0%
7	11.7%	21.3%	44.3%	20.9%	1.7%	0.1%	100.0%
8	14.6%	21.5%	39.3%	22.1%	2.3%	0.1%	100.0%
9	18.5%	20.1%	40.7%	18.4%	2.2%	0.1%	100.0%
10	14.0%	20.2%	43.3%	20.3%	2.2%	0.1%	100.0%
11	12.6%	20.7%	45.0%	19.6%	2.1%	0.0%	100.0%
12	13.8%	24.4%	44.4%	16.1%	1.3%	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, CAL conducted a Writing scaling study. Detailed information about the derivation of the Writing rating scale as well as the psychometric properties of the Writing rating scale are available in the 2016 scaling report (Center for Applied Linguistics, 2017). In 2019–2020, we redesigned the Writing test to allow for embedded field testing, reducing the number of operational tasks from three to two. For details on how we retained the 2016 rating scale parameters and maintained the Writing score scale, see Center for Applied Linguistics (2019).

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

Scale Scores and Proficiency Level Scores

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

- L: (Ability Measure in Logits * 37.571) + 316.637
- R: (Ability Measure in Logits * 26.000) + 323.272
- W: (Ability Measure in Logits * 26.851) + 303.332
- S: (Ability Measure in Logits * 29.248) + 265.076

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

PL scores are interpretations of these scale scores in terms of the proficiency levels described in the WIDA ELD Standards. These interpretations derive from a series of standard-setting studies, in which educators reviewed evidence from the test, either in the form of items for the selected response sections (Listening and Reading) or student portfolios for the constructed response sections (Writing and Speaking), to establish cut scores between the proficiency levels. The first standard-setting study for ACCESS took place in 2005; it established cut scores for all four domains by grade-level cluster (Kenyon, 2006). The second cut score study took place in 2007; it established cut scores for all four domains by grade level (Kenyon, Ryu, & MacGregor, 2013). These cut scores were used to derive proficiency level scores through the 2015–2016 administration (Series 400) of ACCESS for ELLs. WIDA and CAL conducted a third cut score study in summer 2016 (Cook & MacGregor, 2017). The purpose of this study was to re-examine cut scores for each of the proficiency levels in light of the migration from the paper-and-pencil–only assessment to both online and paper delivery, the revision of the Speaking test, and the influence of college- and career-ready standards. These new cut scores were first used for ACCESS Series 401 (2016–2017 school year).

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

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

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

2.1 Complete Item or Task Analysis and Summary

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

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

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

All Rasch analyses were conducted using the Rasch measurement software program *Winsteps* (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 statistics. 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 of these 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 degrade[1] the measurement system."
- Values between 1.5 and 2.0 are "unproductive for construction of measurement, but not degrading."
- Values between 0.5 and 1.5 should be considered "productive for measurement."
- Values below 0.5 are "less productive for measurement, but not degrading."

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

Because we 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 of 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

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

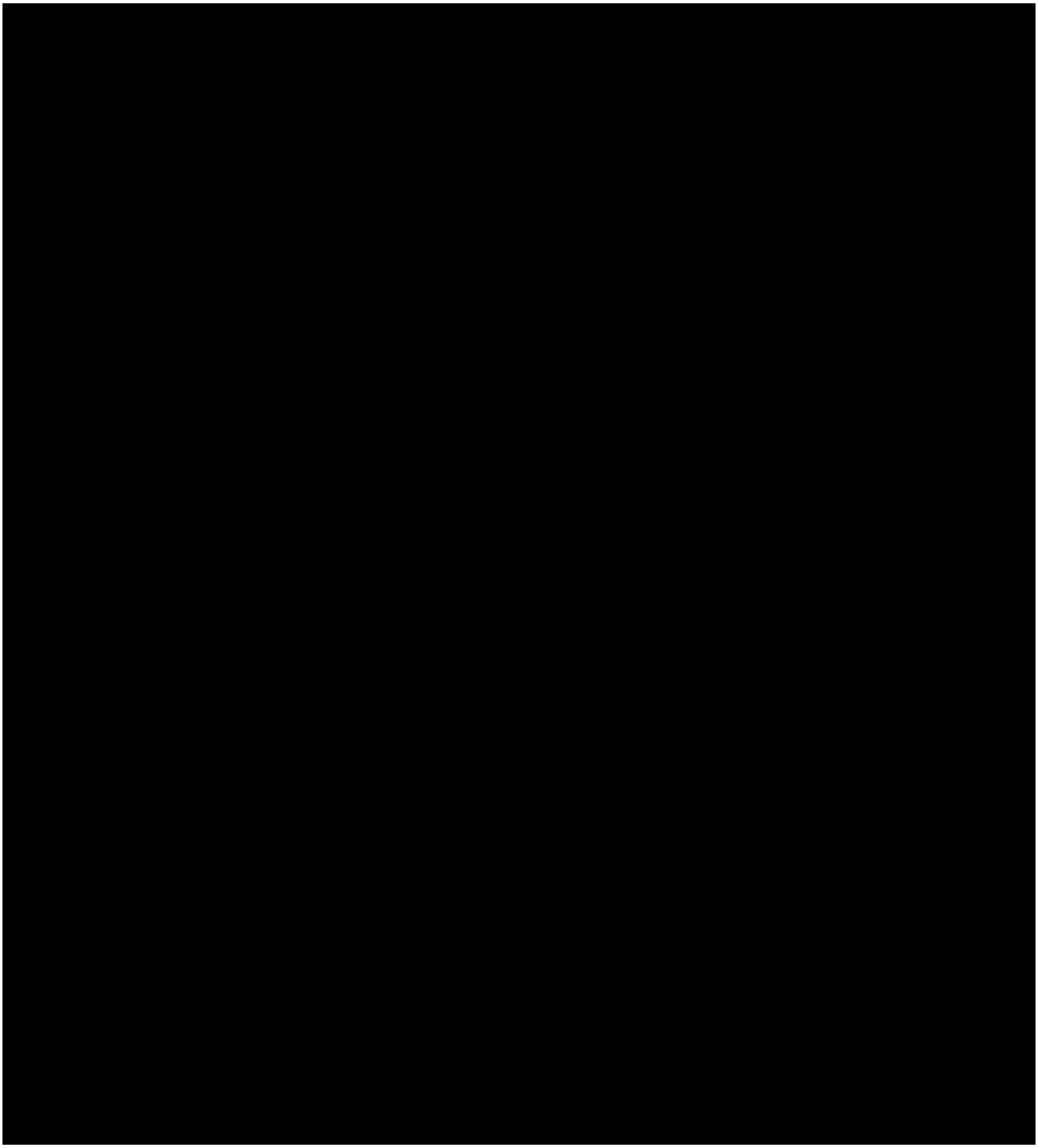
and Reading), the fourth column shows the *p* value (percentage of correct answers on that item). The final two columns show the Rasch fit statistics for the item or task. Folders with items that have fit statistics greater than 2.0 are evaluated by the test development team to determine whether and when the folders can be refreshed in the next test refreshment cycle.

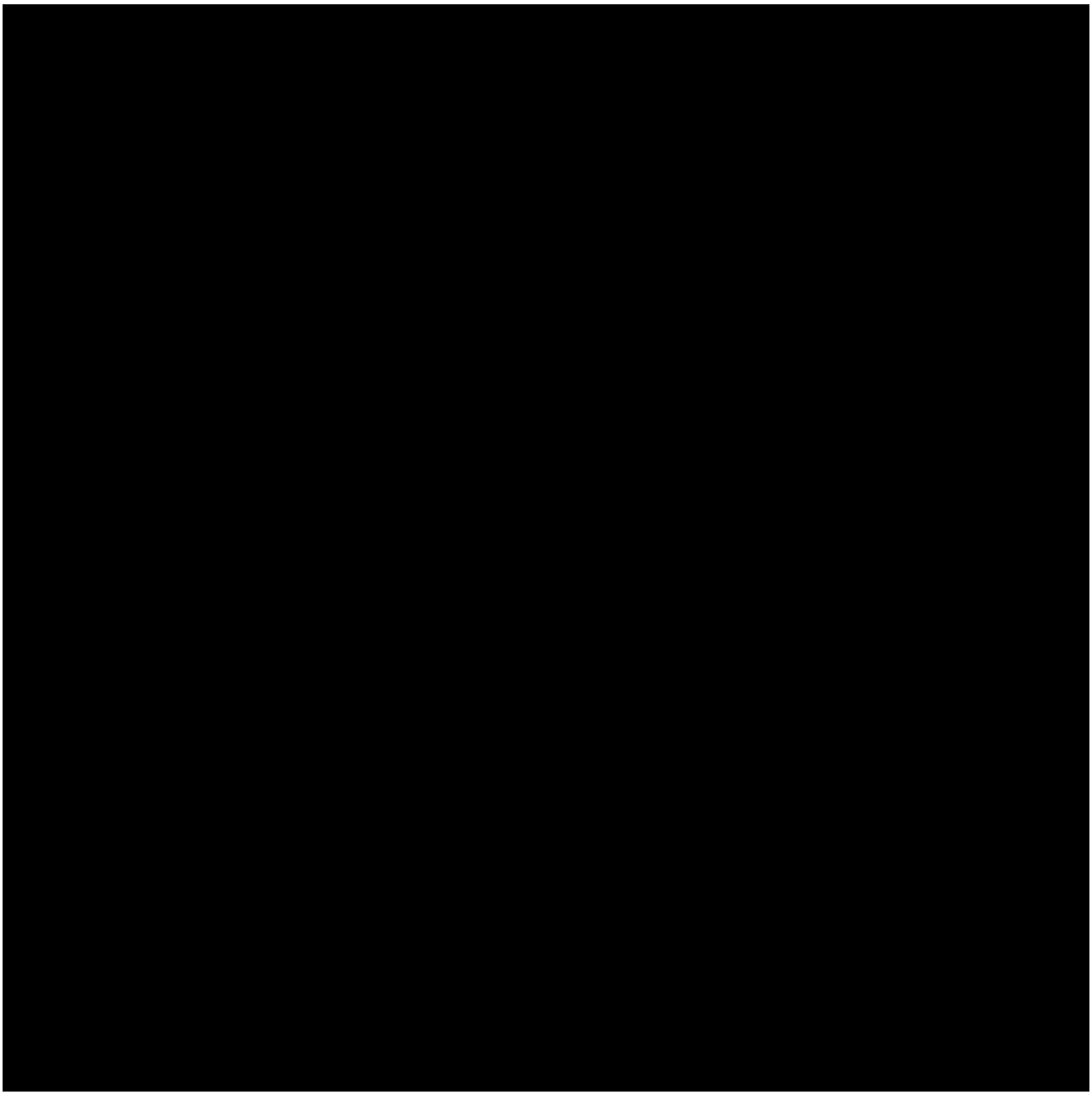
In addition, Writing and Speaking tables have a section at the bottom of the table that provides raw score distributions by task.

The results show that all items and tasks have infit mean square statistics less than 2 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 is one item in Listening grade-level cluster 1, two items in Listening grade-level cluster 2-3, and one item in Listening grade-level cluster 4-5 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.

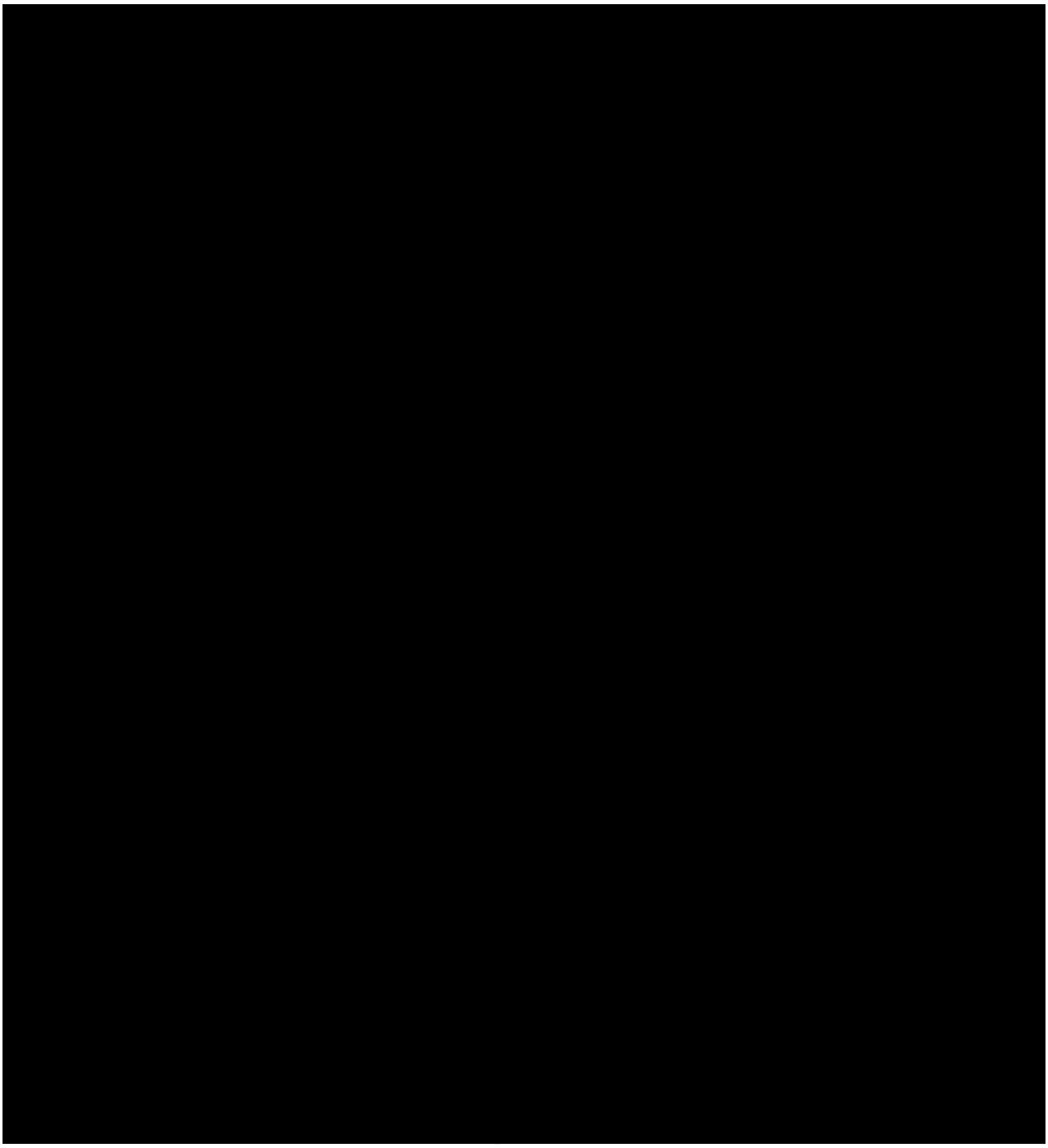
2.1.1 **Listening**

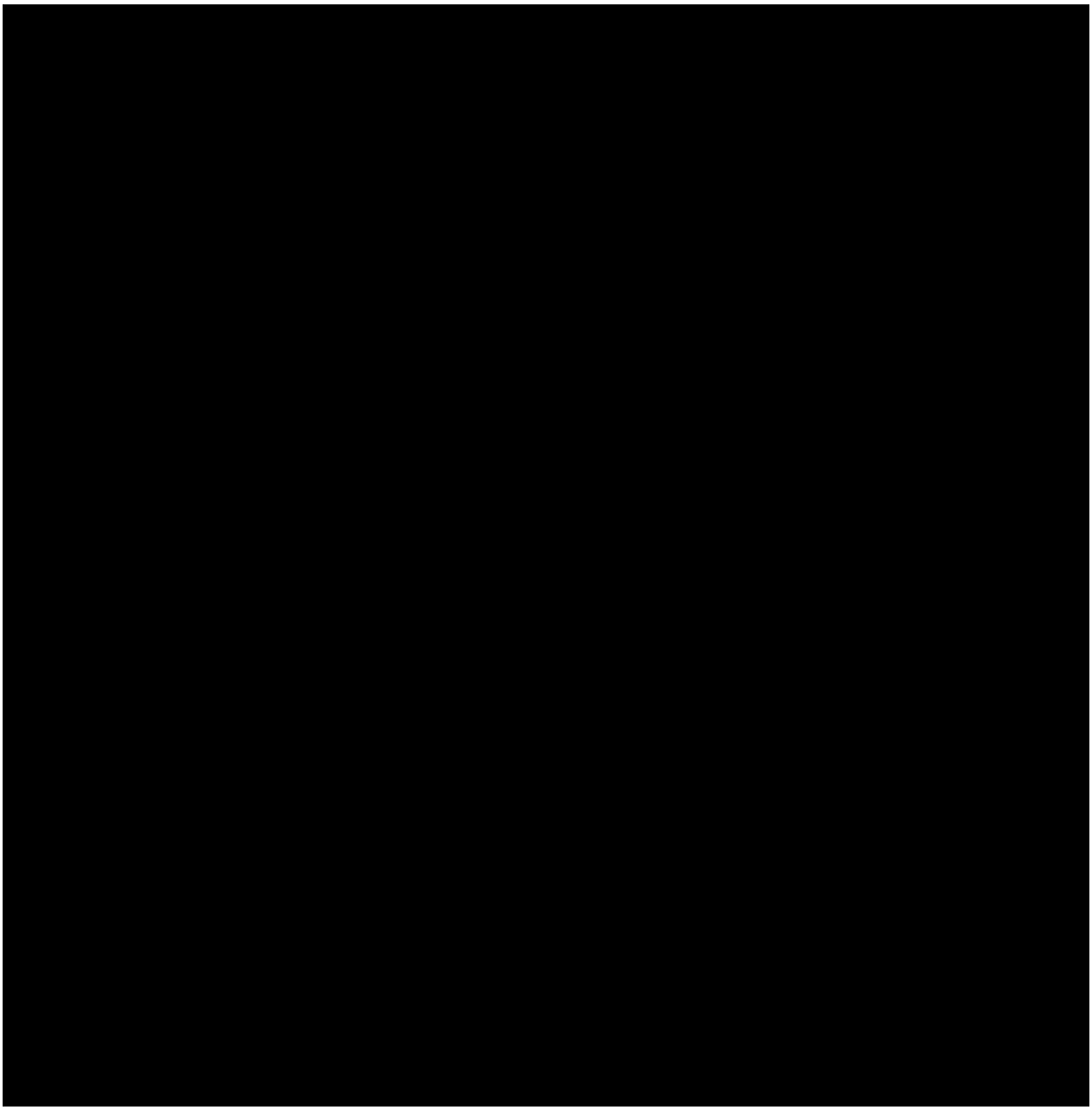
2.1.1.1 *Grade 1*



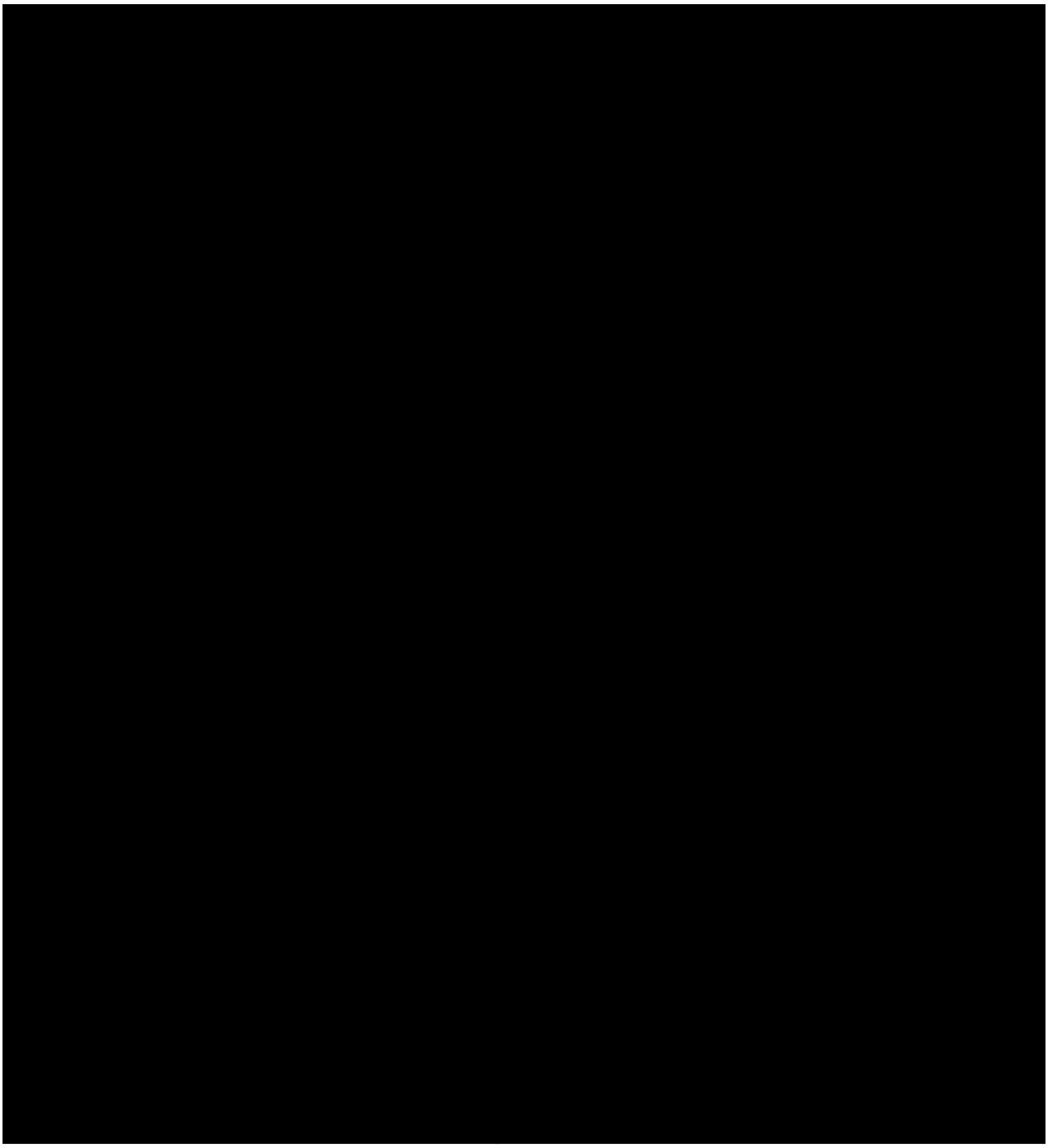


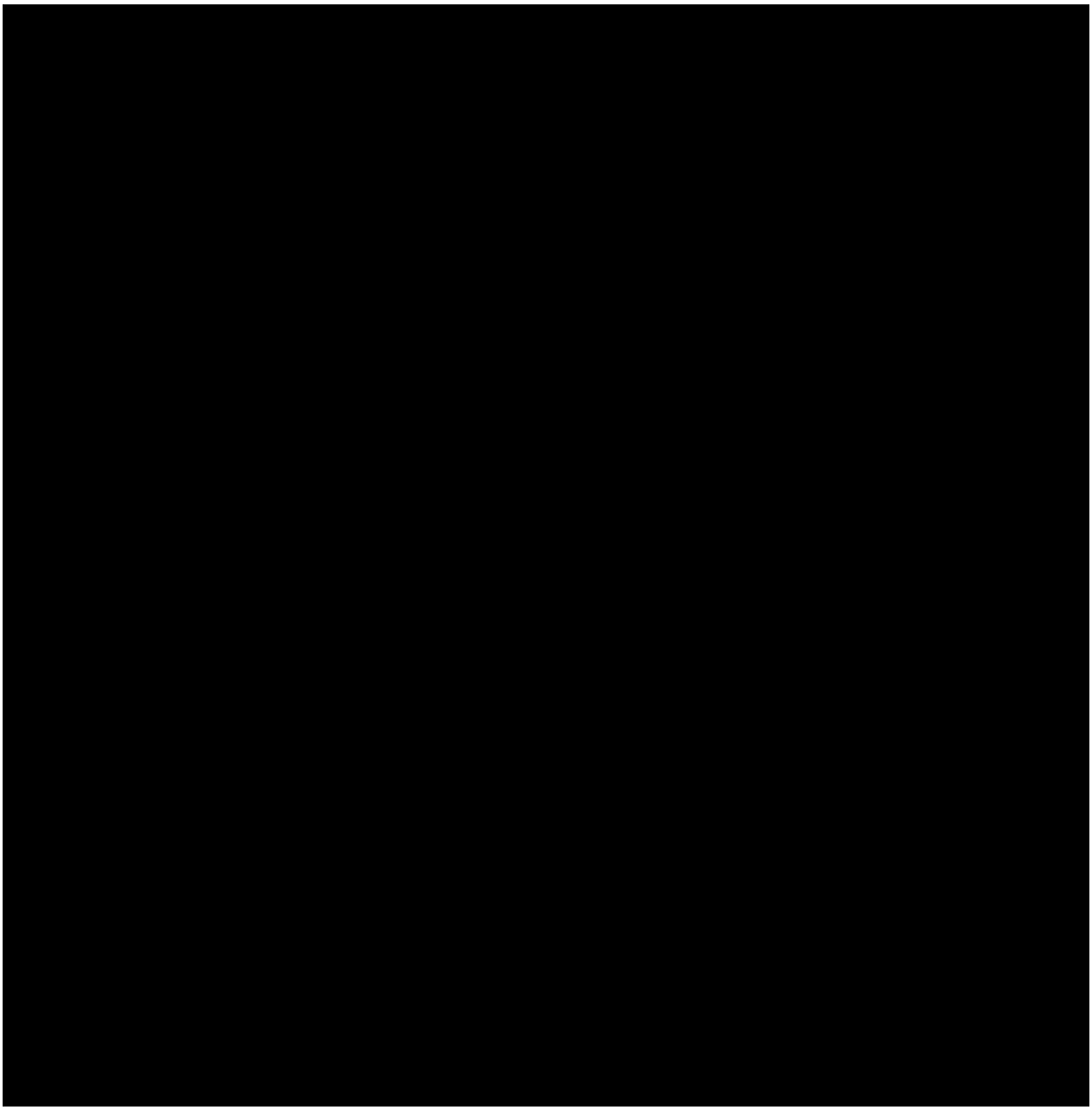
2.1.1.2 *Grades 2–3*



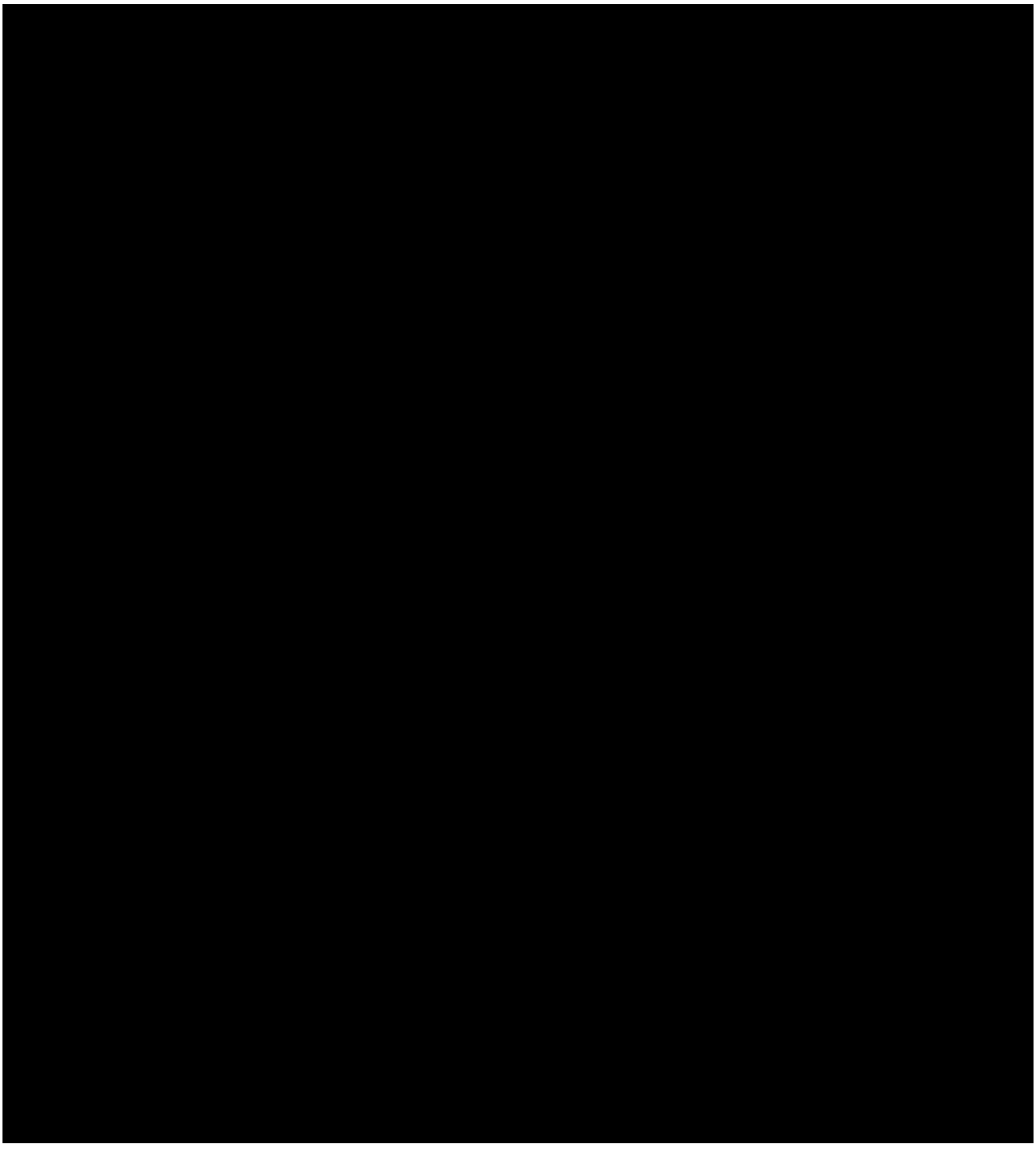


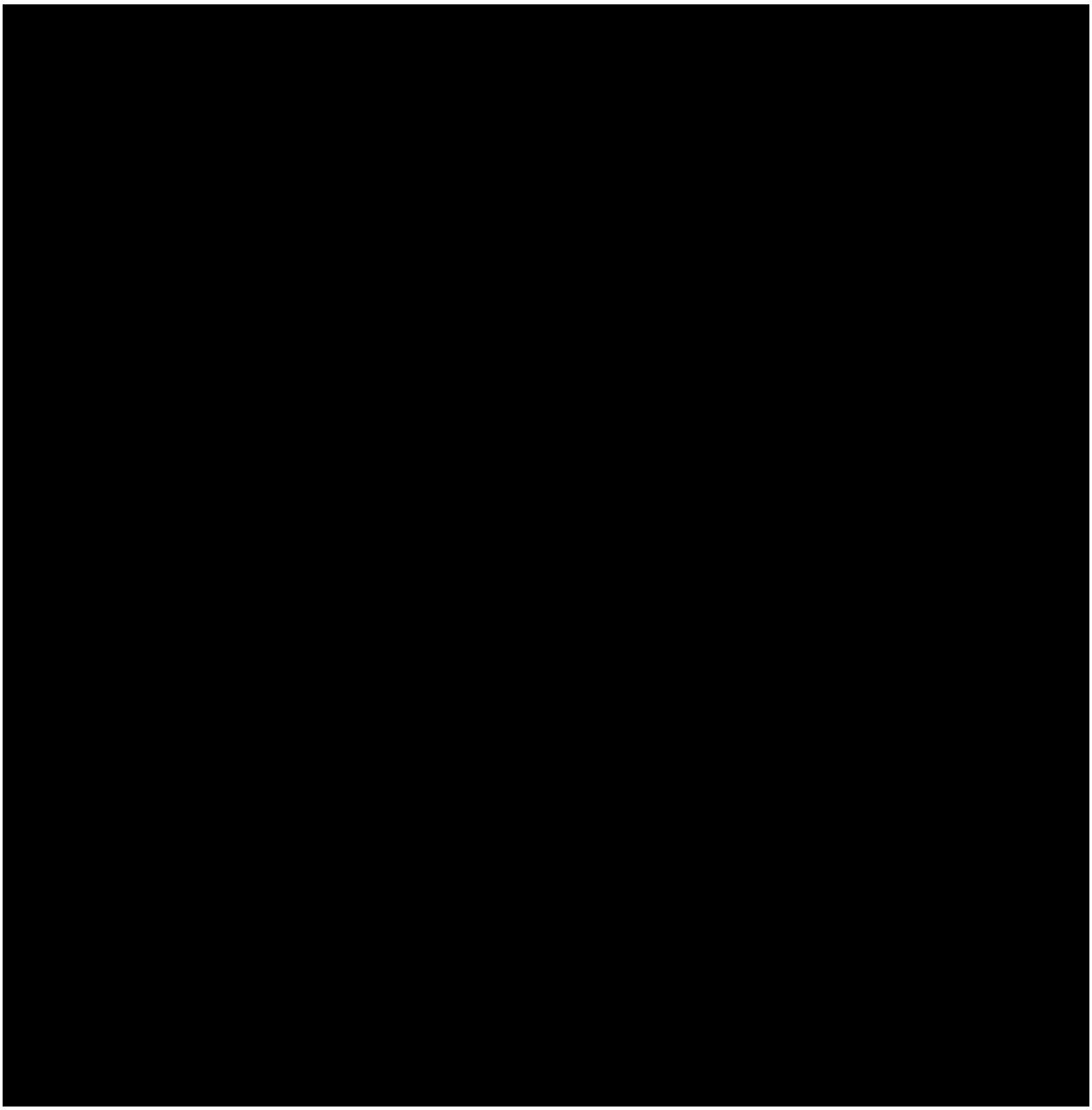
2.1.1.3 *Grades 4–5*



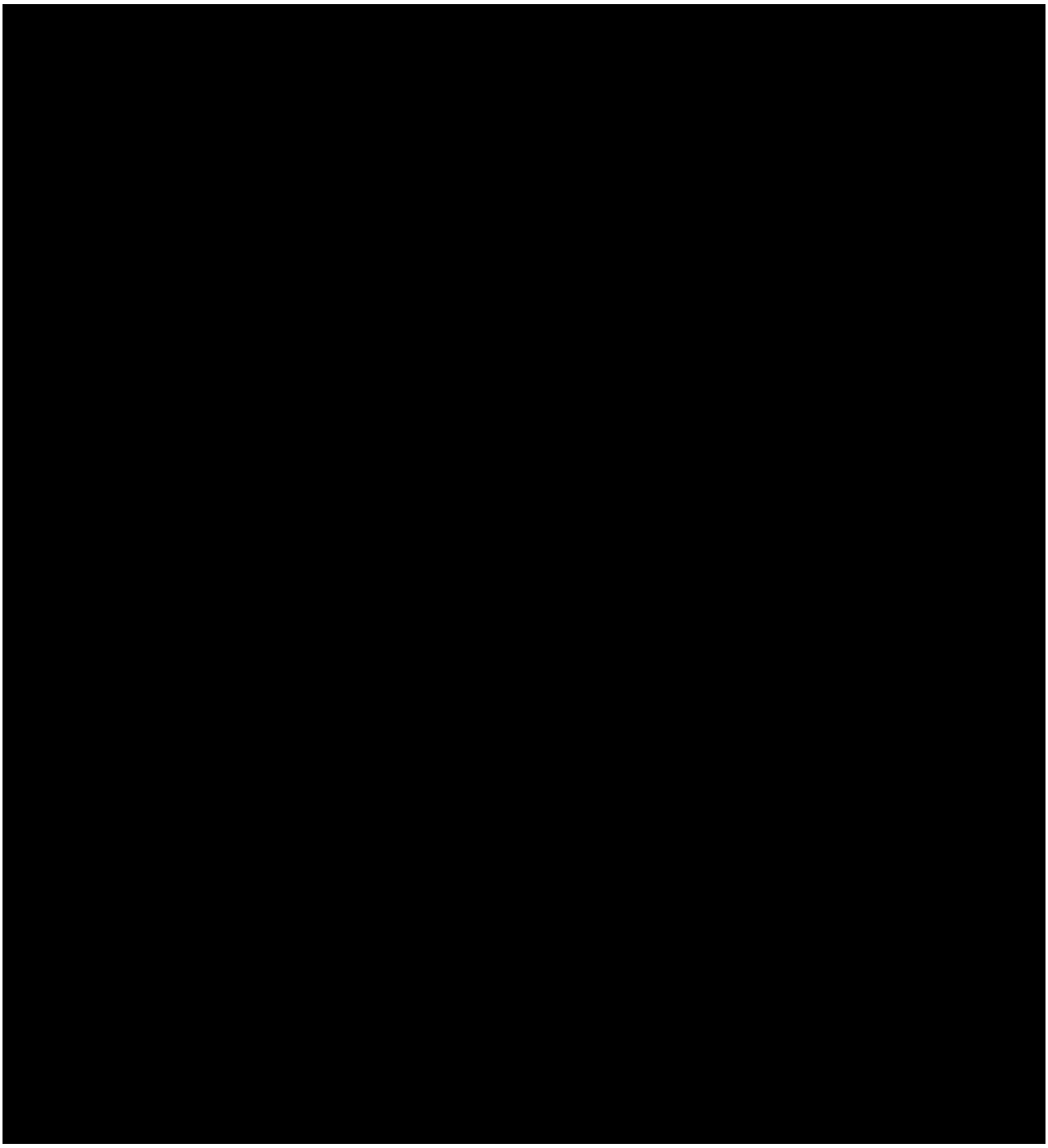


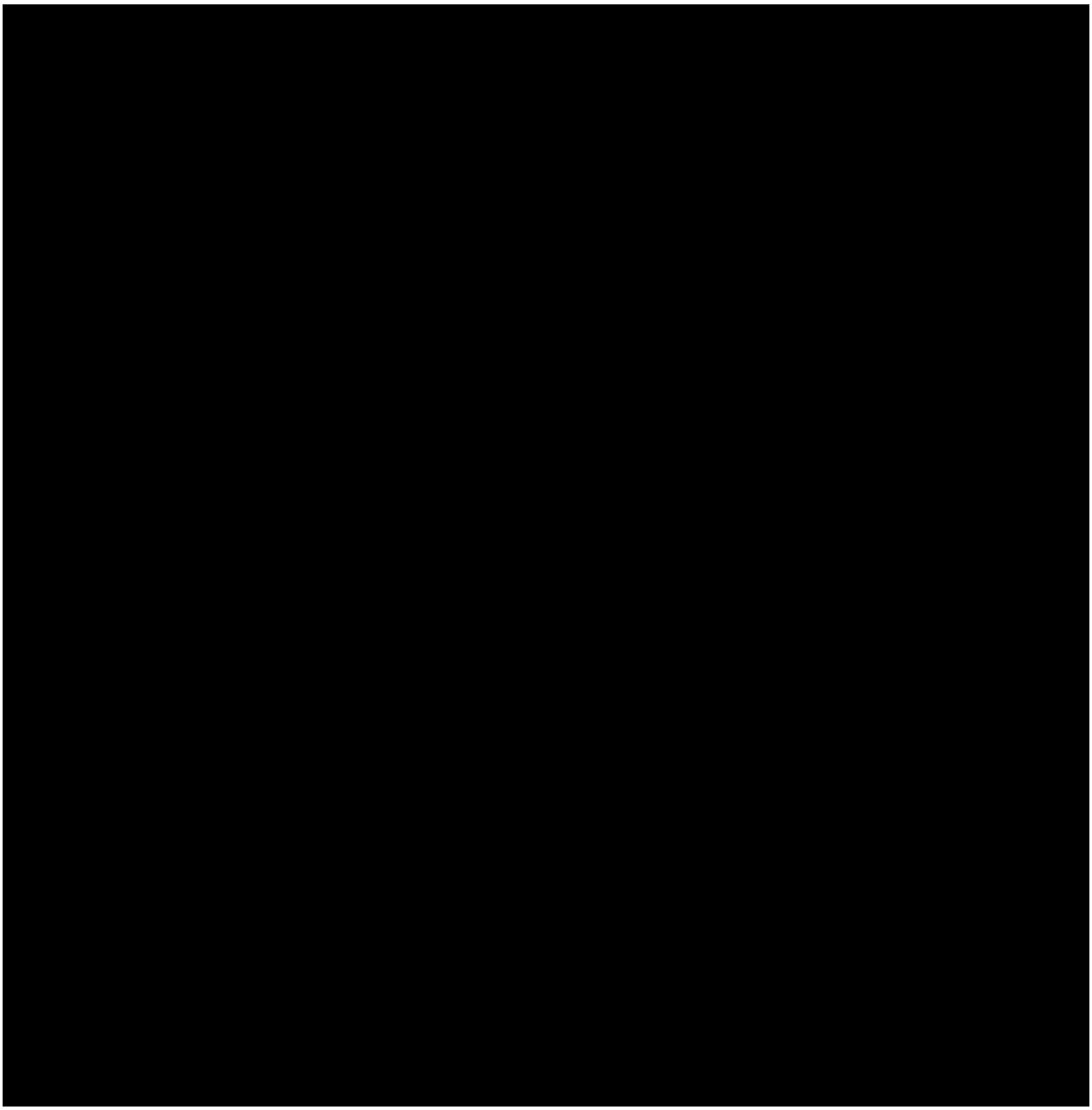
2.1.1.4 *Grades 6–8*





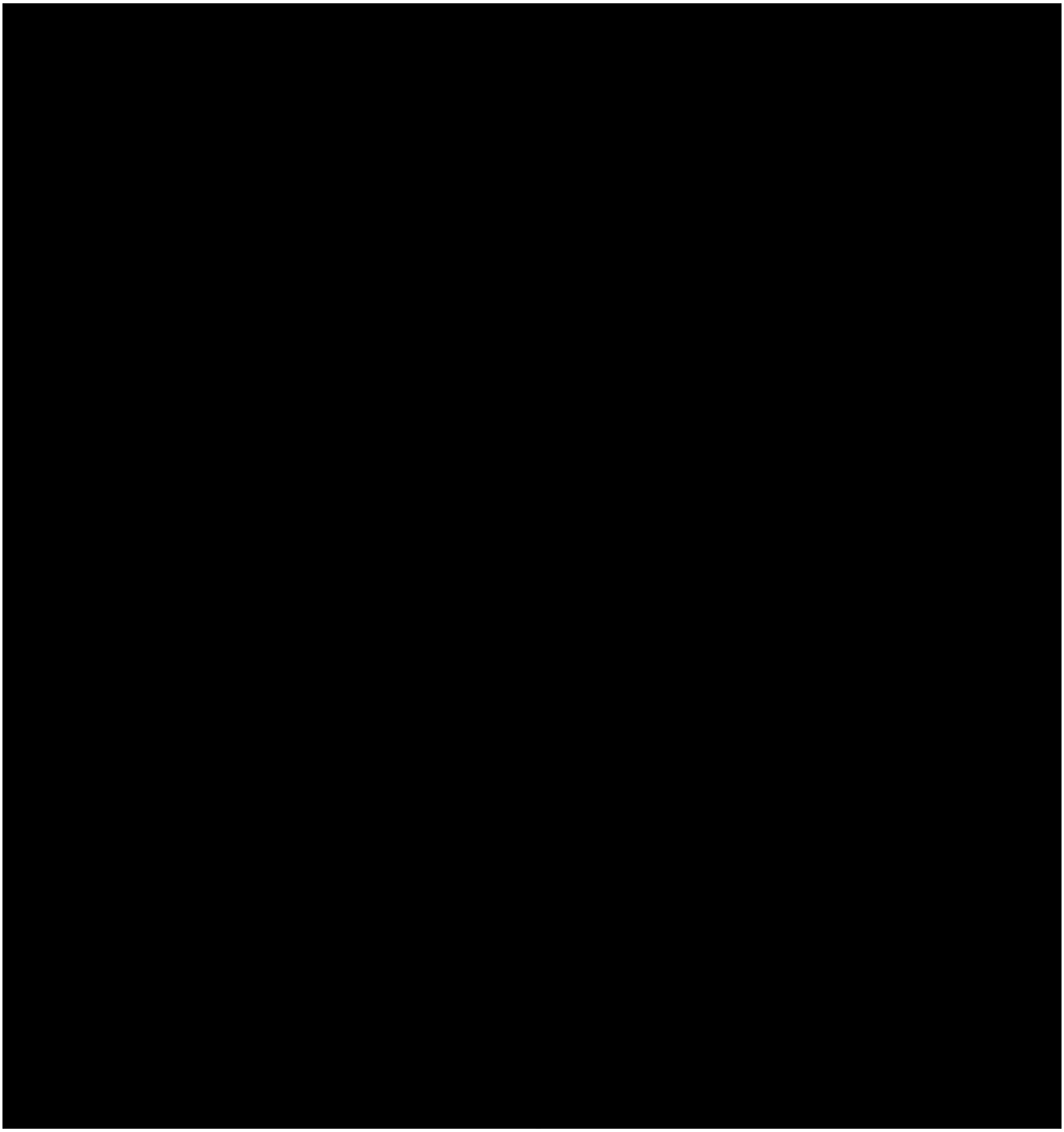
2.1.1.5 *Grades 9–12*

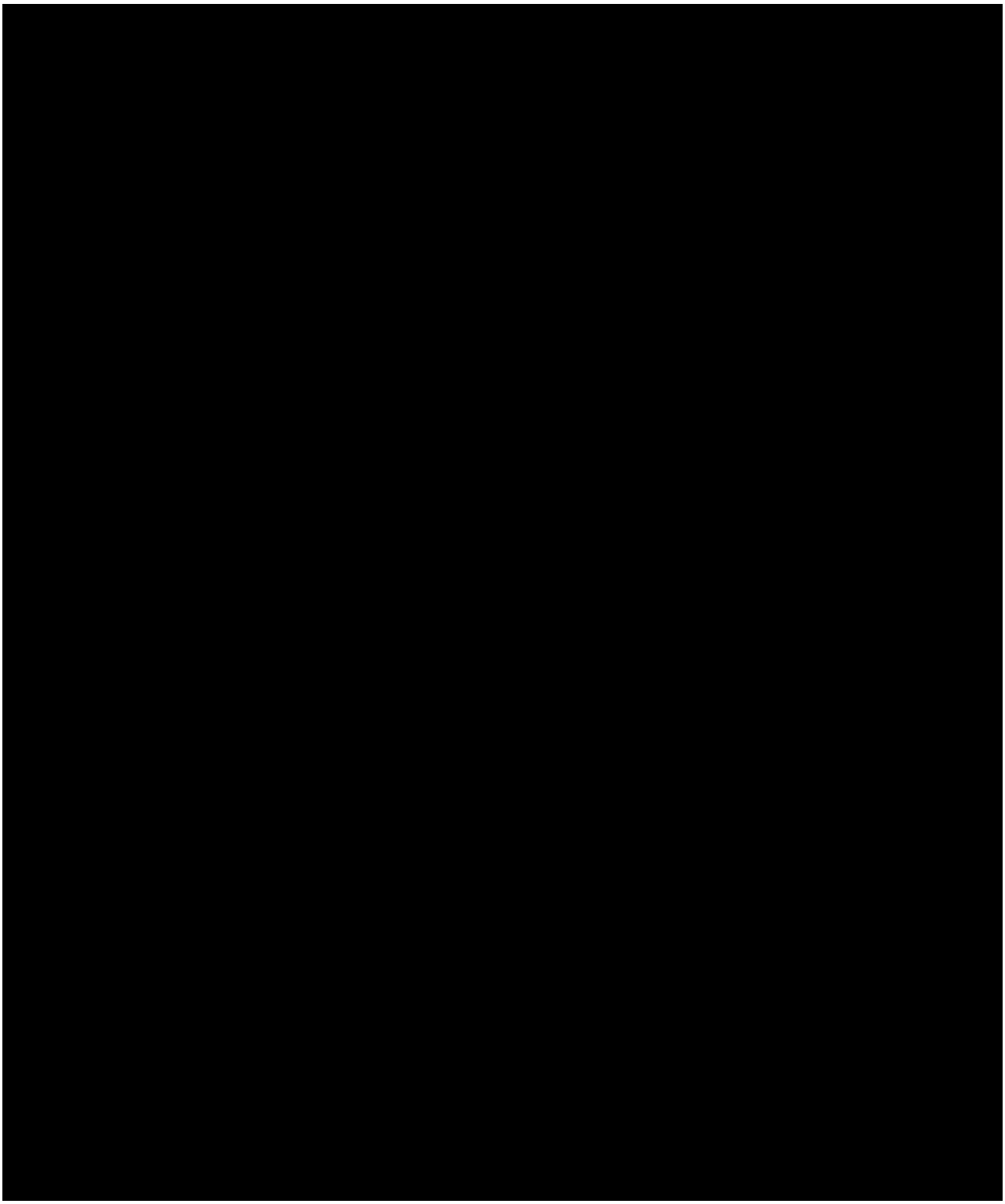




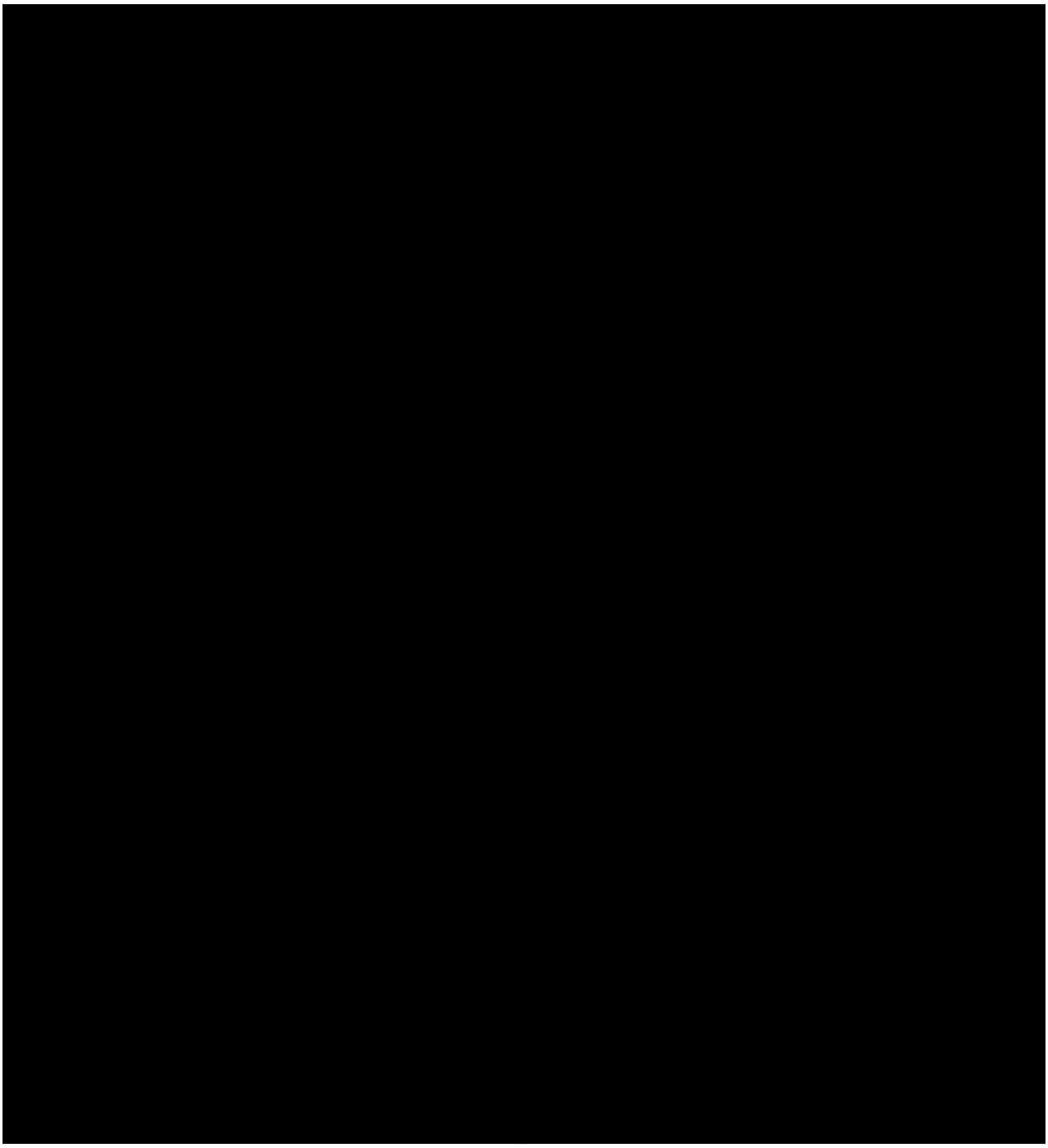
2.1.2 **Reading**

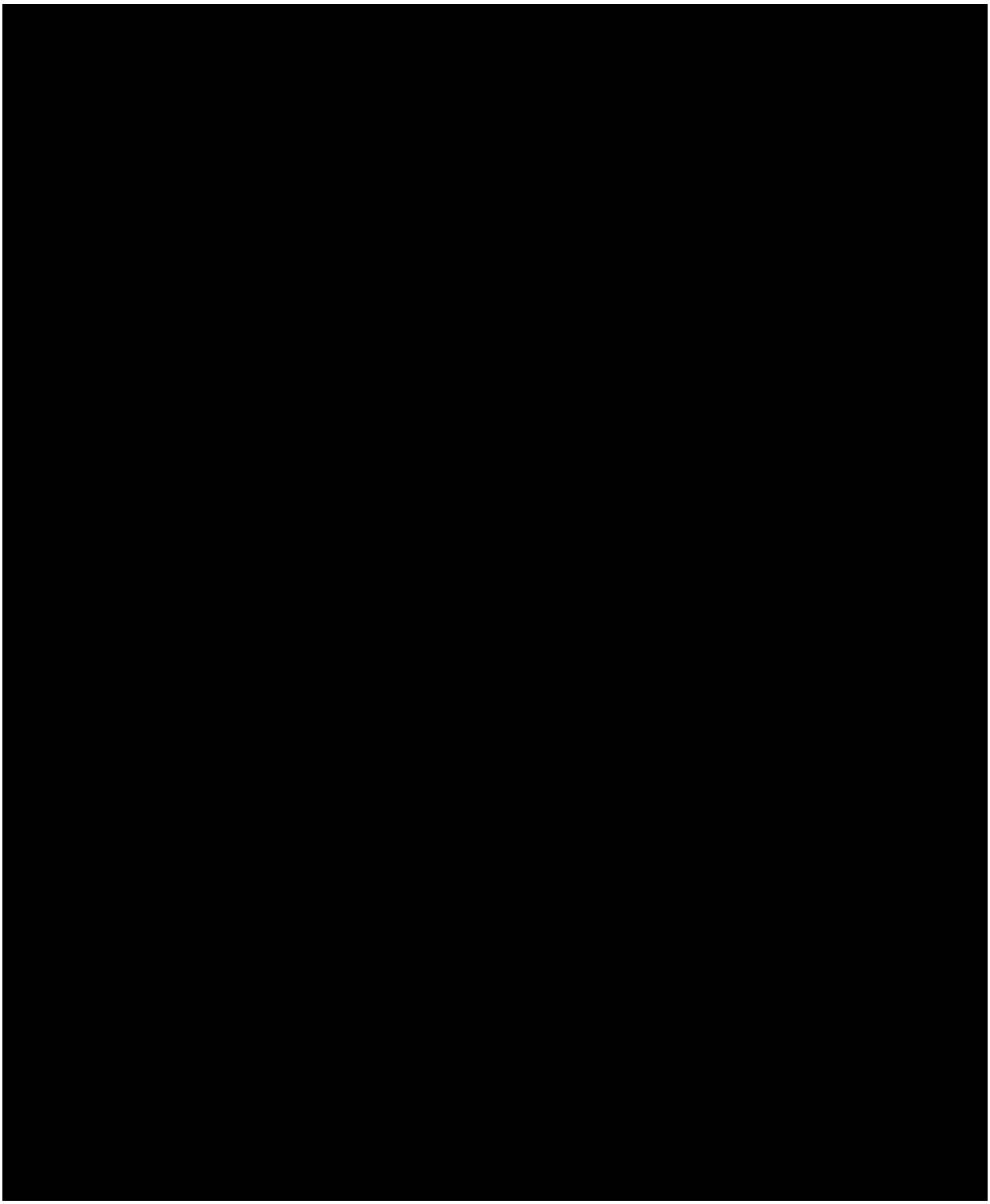
2.1.2.1 *Grade 1*



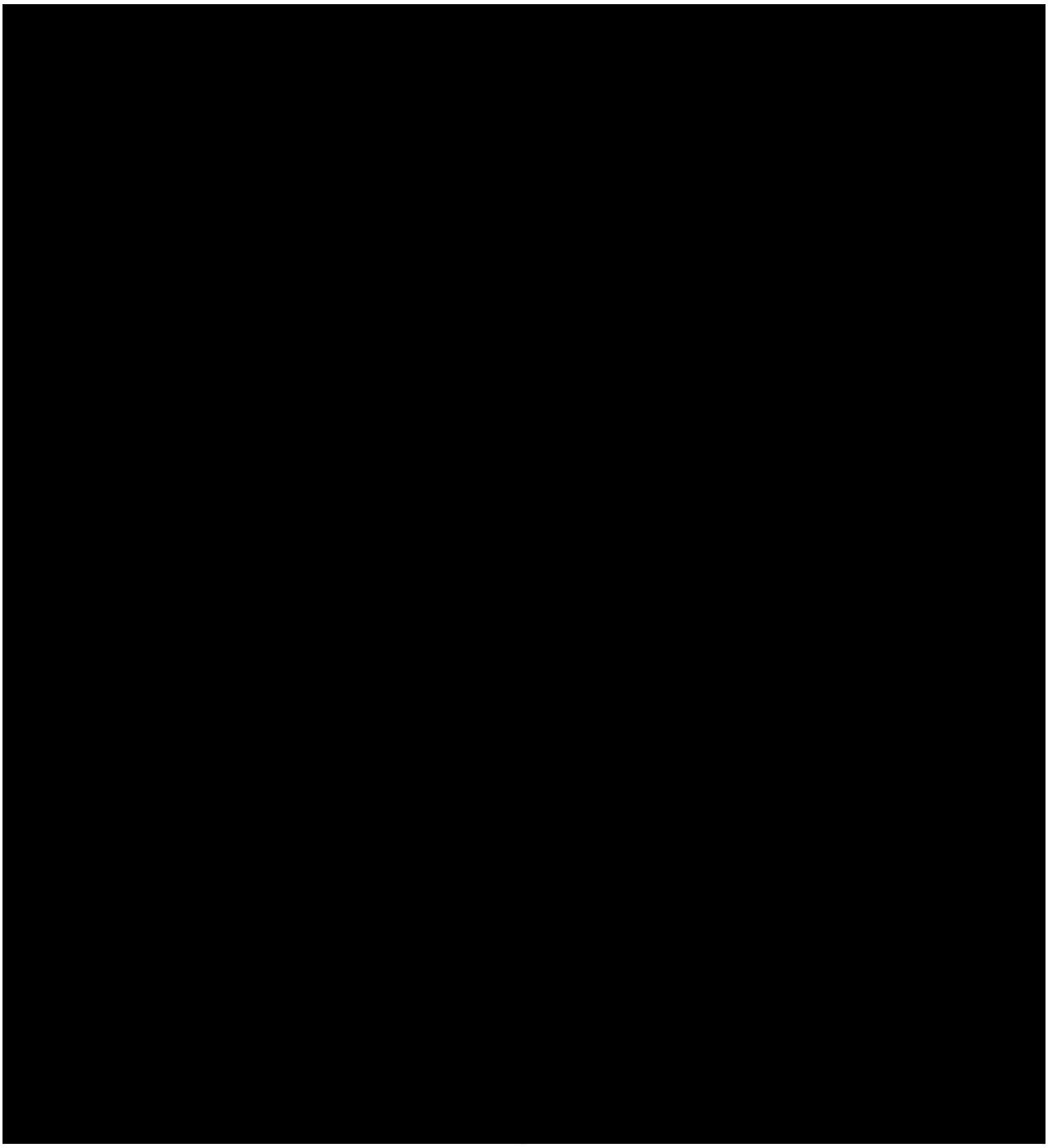


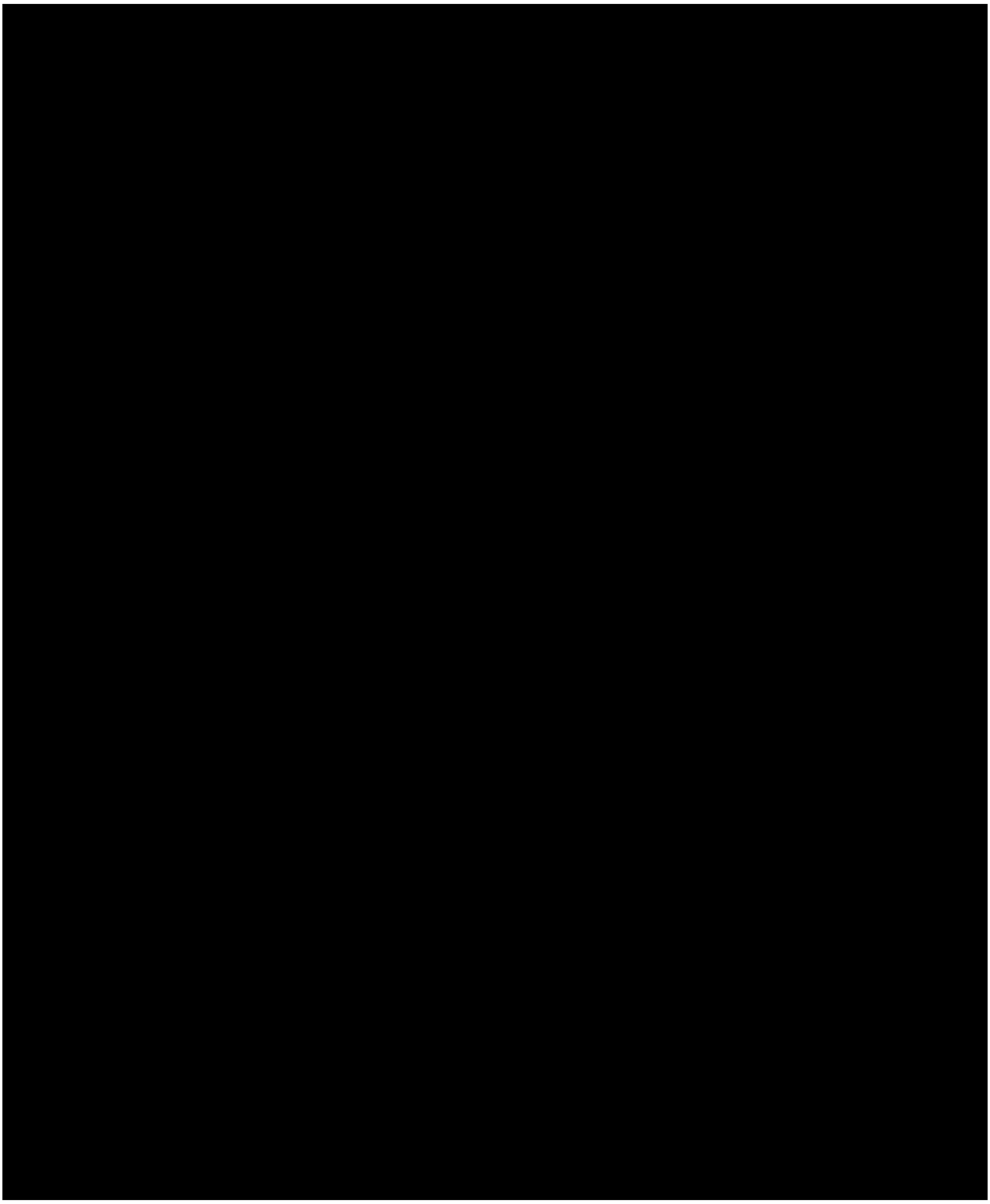
2.1.2.2 *Grades 2–3*



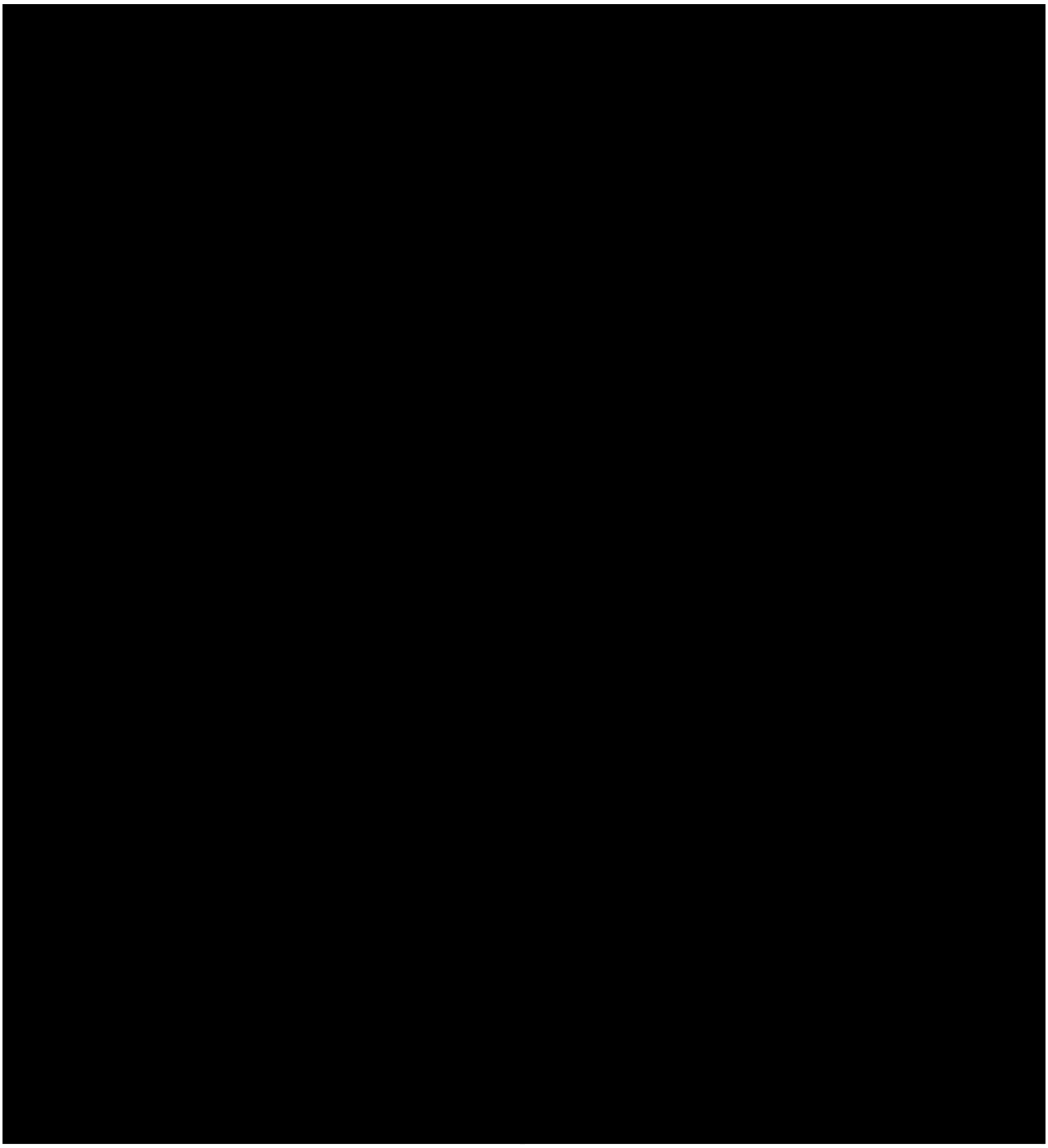


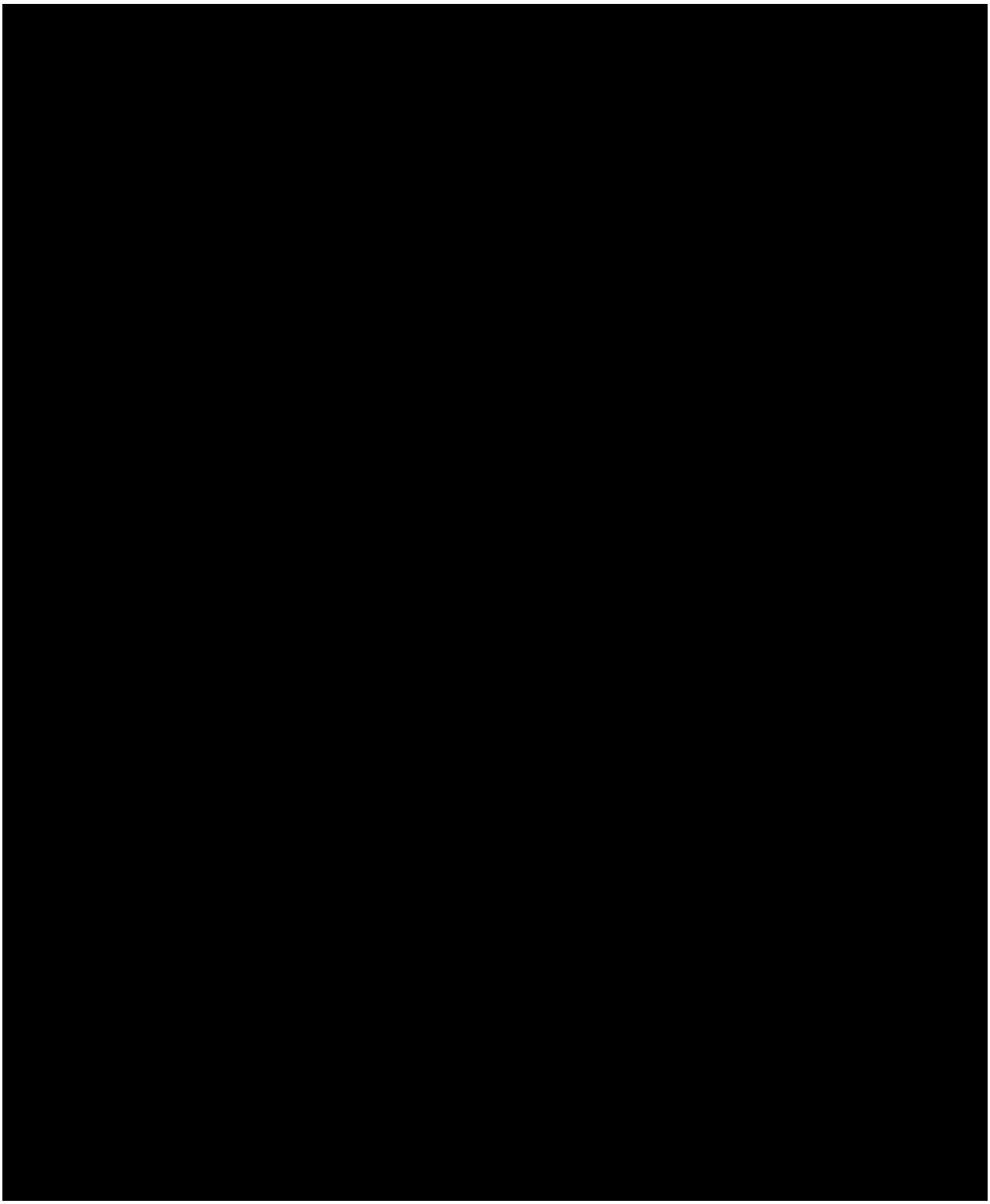
2.1.2.3 *Grades 4–5*



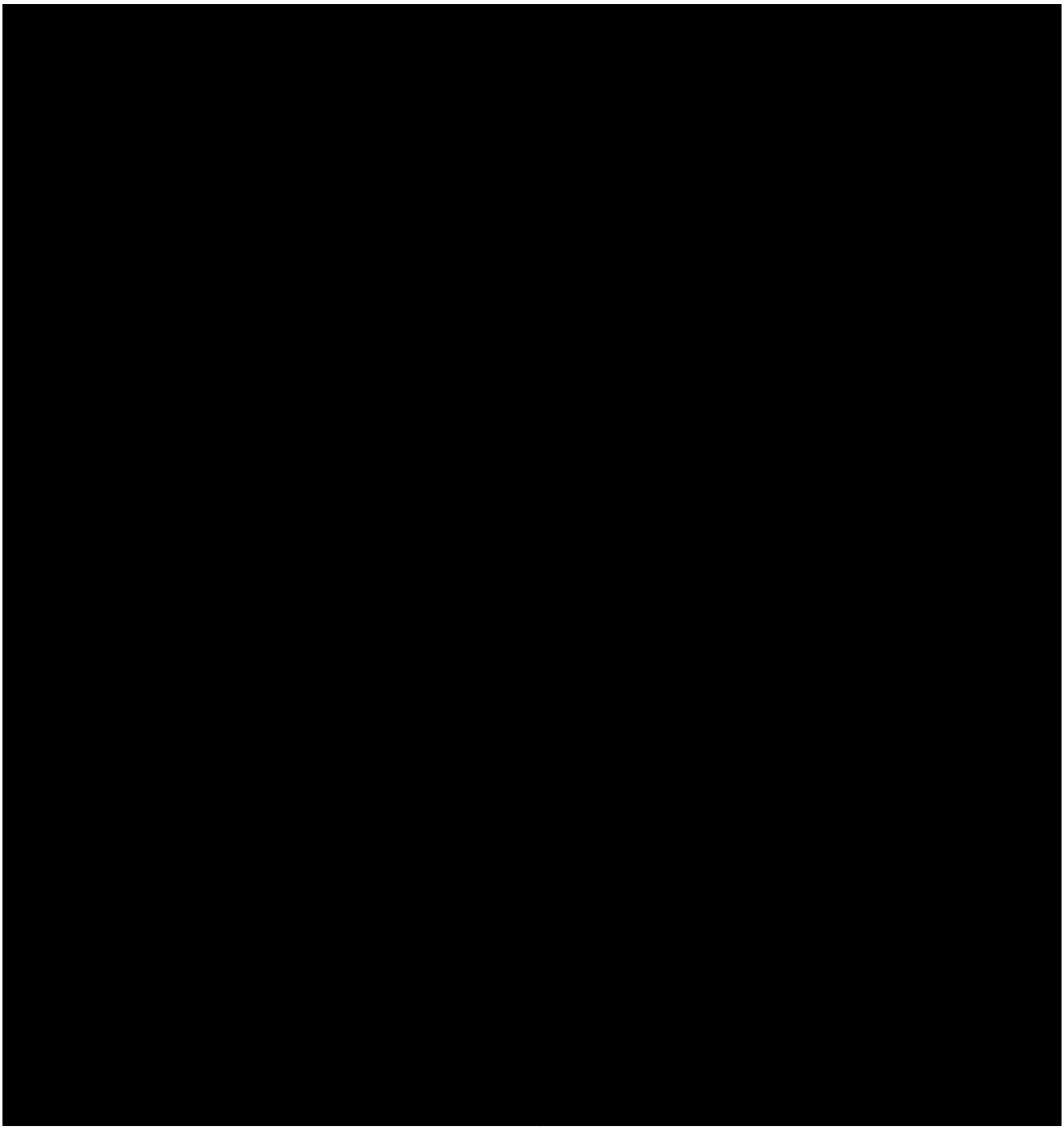


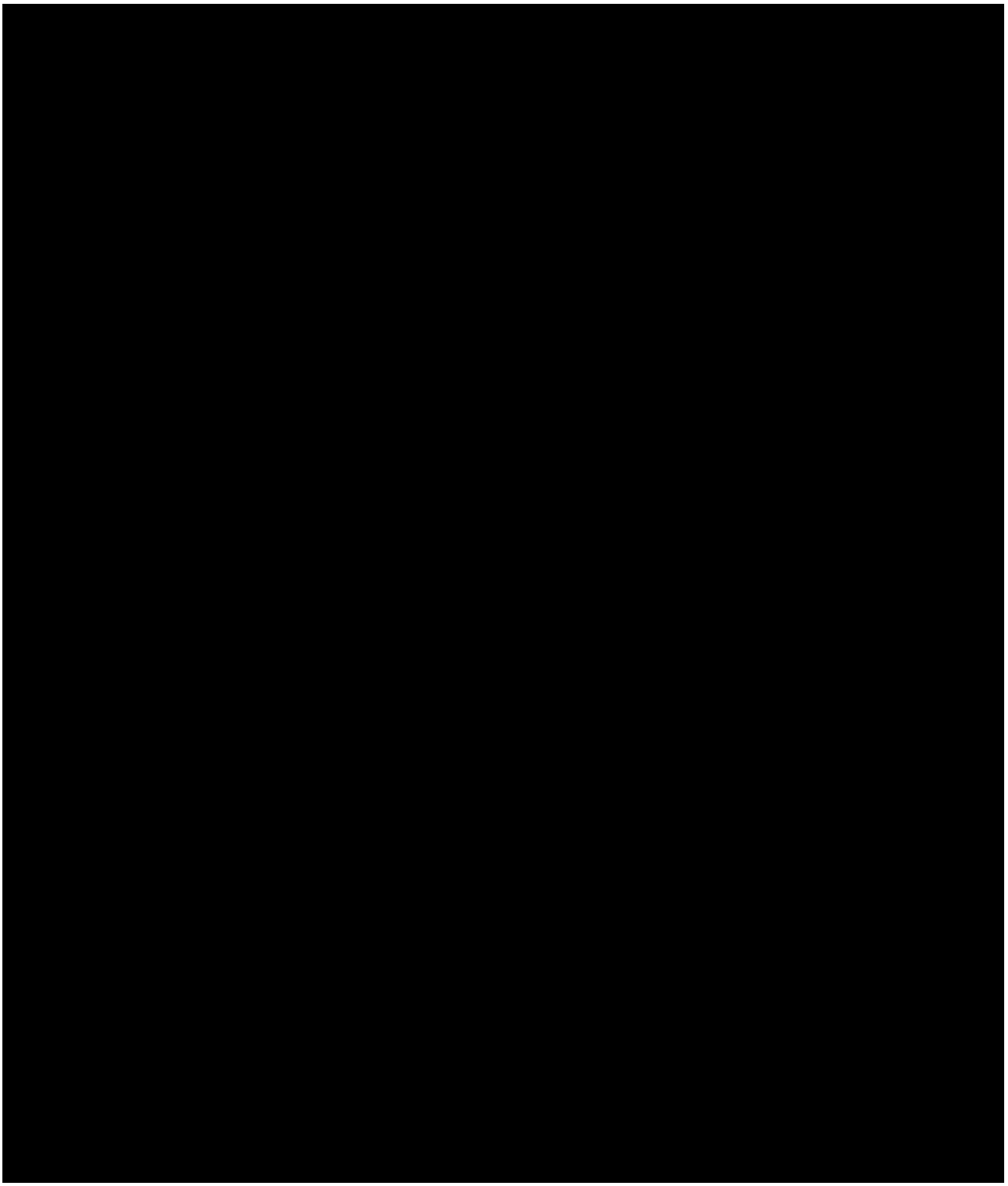
2.1.2.4 *Grades 6–8*





2.1.2.5 *Grades 9–12*





2.1.3 **Writing**

2.1.3.1 *Grade 1*

2.1.3.2 *Grades 2–3*

2.1.3.3 *Grades 4–5*

2.1.3.4 *Grades 6–8*

2.1.3.5 *Grades 9–12*

2.1.4 **Speaking**

2.1.4.1 *Grade 1*

2.1.4.2 *Grades 2–3*

2.1.4.3 *Grades 4–5*

2.1.4.4 *Grades 6–8*

2.1.4.5 *Grades 9–12*

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 ethnicity[2] was unknown from both analyses. We used two commonly used procedures for detecting DIF: one for dichotomously scored items (Listening and Reading), conducted prior to operational testing, and one for polytomously scored items (Writing and Speaking), conducted on population data 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 exactly 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 exactly 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

[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.

students as the matching variable for the M-H DIF procedure. Once we have matched students on the expected true score, the ordinary M-H DIF procedure and the ETS evaluation criterion for severity of M-H DIF can be applied. In CAL's implementation of this method, students are matched for M-H DIF analysis on the basis of 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 2.2.1). 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 English learners (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.

For Speaking and Writing tasks, there is not sufficient scored data for DIF analysis of these 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 501, one item in Listening Grades 2–3 showed C-level DIF. The item was reviewed by a panel as described above. The panel was not able to detect any reason for bias in the performance of this item and recommended that the item be retained on the assessment.

2.2.1 **Listening**

2.2.1.1 *Grade 1*

2.2.1.2 *Grades 2–3*

2.2.1.3 *Grades 4–5*

2.2.1.4 *Grades 6–8*

2.2.1.5 *Grades 9–12*

2.2.2 **Reading**

2.2.2.1 *Grade 1*

2.2.2.2 *Grades 2–3*

2.2.2.3 *Grades 4–5*

2.2.2.4 *Grades 6–8*

2.2.2.5 *Grades 9–12*

2.2.3 Writing

2.2.3.1 *Grade 1*

Table 2.2.3.1.1

DIF Analysis and Summary: Writ 1 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.3.1.2

DIF Analysis and Summary: Writ 1 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

2.2.3.2 *Grades 2–3*

Table 2.2.3.2.1

DIF Analysis and Summary: Writ 2-3 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.3.2.2

DIF Analysis and Summary: Writ 2-3 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

2.2.3.3 *Grades 4–5*

Table 2.2.3.3.1

DIF Analysis and Summary: Writ 4-5 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.3.3.2

DIF Analysis and Summary: Writ 4-5 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

2.2.3.4 *Grades 6–8*

Table 2.2.3.4.1

DIF Analysis and Summary: Writ 6-8 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.3.4.2

DIF Analysis and Summary: Writ 6-8 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

2.2.3.5 *Grades 9–12*

Table 2.2.3.5.1

DIF Analysis and Summary: Writ 9-12 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.3.5.2

DIF Analysis and Summary: Writ 9-12 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	1	1	1
BB	0	0	0	0
CC	0	0	0	0

2.2.4 **Speaking**

2.2.4.1 *Grade 1*

Table 2.2.4.1.1

DIF Analysis and Summary: Spek 1 Pre-A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	**Favoring Male (M)**	**Favoring Female (F)**	**Favoring Hispanic (H)**	**Favoring Other (O)**
AA	2	1	2	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.1.2

DIF Analysis and Summary: Spek 1 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	**Favoring Male (M)**	**Favoring Female (F)**	**Favoring Hispanic (H)**	**Favoring Other (O)**
AA	4	2	4	2
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.1.3

DIF Analysis and Summary: Spek 1 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	**Favoring Male (M)**	**Favoring Female (F)**	**Favoring Hispanic (H)**	**Favoring Other (O)**
AA	3	3	4	2
BB	0	0	0	0
CC	0	0	0	0

2.2.4.2 *Grades 2–3*

Table 2.2.4.2.1

DIF Analysis and Summary: Spek 2-3 Pre-A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	2	2	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.2.2

DIF Analysis and Summary: Spek 2-3 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	3	3	2	4
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.2.3

DIF Analysis and Summary: Spek 2-3 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	4	2	2	4
BB	0	0	0	0
CC	0	0	0	0

2.2.4.3 *Grades 4–5*

Table 2.2.4.3.1

DIF Analysis and Summary: Spek 4-5 Pre-A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	2	1	2
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.3.2

DIF Analysis and Summary: Spek 4-5 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	3	3	3	3
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.3.3

DIF Analysis and Summary: Spek 4-5 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	3	3	2	4
BB	0	0	0	0
CC	0	0	0	0

2.2.4.4 *Grades 6–8*

Table 2.2.4.4.1

DIF Analysis and Summary: Spek 6-8 Pre-A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	2	2	1
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.4.2

DIF Analysis and Summary: Spek 6-8 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	2	4	2	4
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.4.3

DIF Analysis and Summary: Spek 6-8 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	3	3	1	5
BB	0	0	0	0
CC	0	0	0	0

2.2.4.5 *Grades 9–12*

Table 2.2.4.5.1

DIF Analysis and Summary: Spek 9-12 Pre-A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	1	2	1	2
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.5.2

DIF Analysis and Summary: Spek 9-12 A S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	2	4	3	3
BB	0	0	0	0
CC	0	0	0	0

Table 2.2.4.5.3

DIF Analysis and Summary: Spek 9-12 B/C S501 Online

DIF Summary	Male/Female		Hispanic/Other	
DIF Level	Favoring Male (M)	Favoring Female (F)	Favoring Hispanic (H)	Favoring Other (O)
AA	3	3	4	2
BB	0	0	0	0
CC	0	0	0	0

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 also 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 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 are able to achieve all 6 points on this form.

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 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	158,459	0	13	5.99	2.66
Total	158,459	0	13	5.99	2.66

Figure 2.3.3.1.1
Raw Scores: Writ 1 A S501 Online

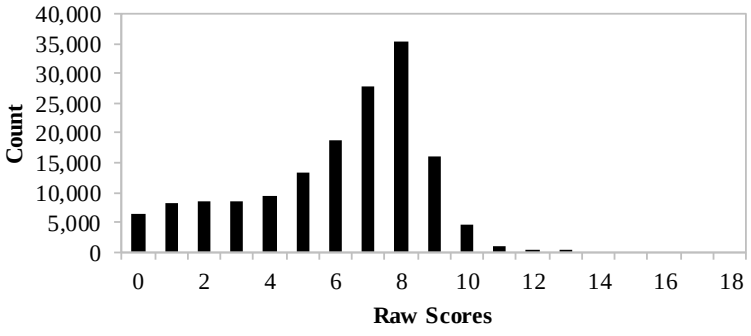
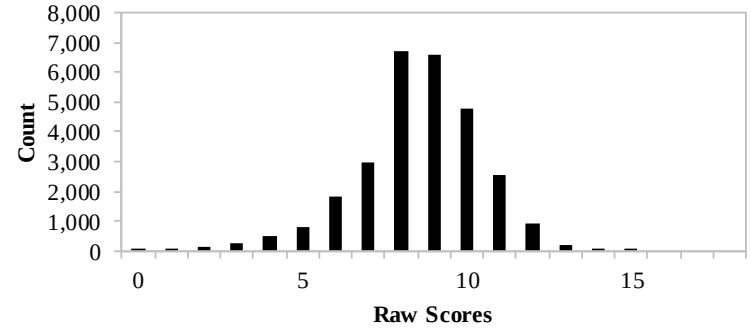


Table 2.3.3.1.2

Raw Score Descriptive Statistics: Writ 1 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	28,391	0	15	8.50	1.94
Total	28,391	0	15	8.50	1.94

Figure 2.3.3.1.2
Raw Scores: Writ 1 B/C S501 Online



2.3.3.2 *Grades 2–3*

Table 2.3.3.2.1

Raw Score Descriptive Statistics: Writ 2-3 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	56,977	0	16	5.94	3.00
3	38,672	0	15	6.45	3.07
Total	95,649	0	16	6.15	3.04

Figure 2.3.3.2.1
Raw Scores: Writ 2-3 A S501 Online

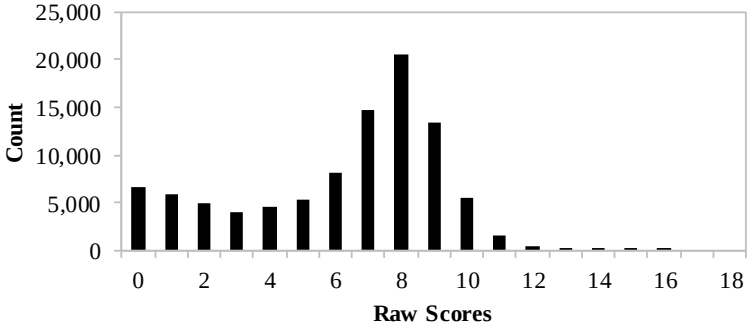
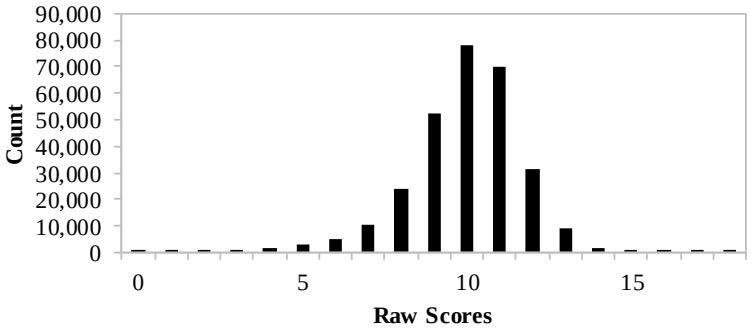


Table 2.3.3.2.2

Raw Score Descriptive Statistics: Writ 2-3 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	137,151	0	16	9.34	1.91
3	153,337	0	18	10.39	1.60
Total	290,488	0	18	9.89	1.83

Figure 2.3.3.2.2
Raw Scores: Writ 2-3 B/C S501 Online



2.3.3.3 *Grades 4–5*

Table 2.3.3.3.1

Raw Score Descriptive Statistics: Writ 4-5 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	24,984	0	14	4.82	3.03
5	24,928	0	15	5.51	3.01
Total	49,912	0	15	5.16	3.04

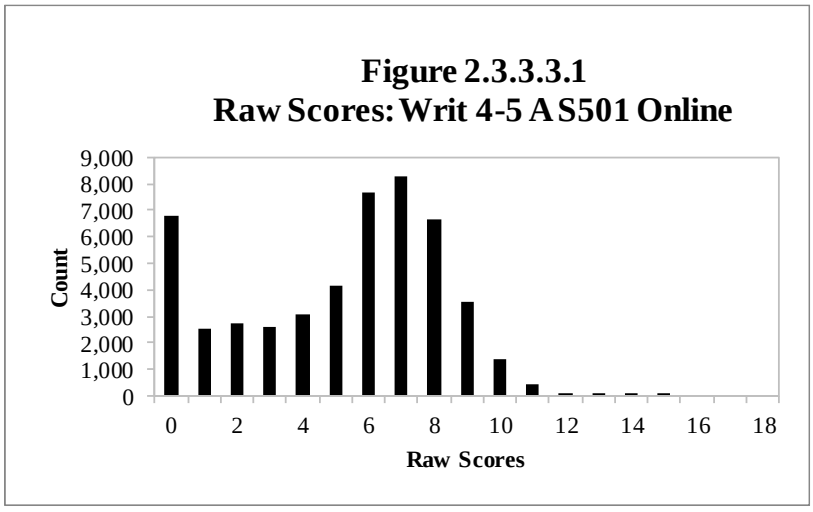
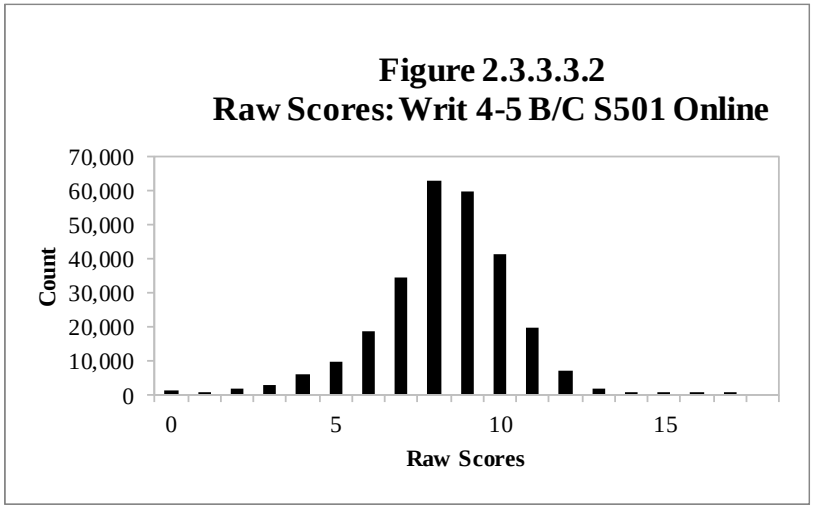


Table 2.3.3.3.2

Raw Score Descriptive Statistics: Writ 4-5 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	150,873	0	17	7.98	2.08
5	117,540	0	17	8.70	1.91
Total	268,413	0	17	8.30	2.04



2.3.3.4 *Grades 6–8*

Table 2.3.3.4.1

Raw Score Descriptive Statistics: Writ 6-8 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	33,634	0	13	5.75	2.53
7	39,214	0	14	6.24	2.52
8	37,263	0	15	6.48	2.52
Total	110,111	0	15	6.17	2.54

Figure 2.3.3.4.1
Raw Scores: Writ 6-8 A S501 Online

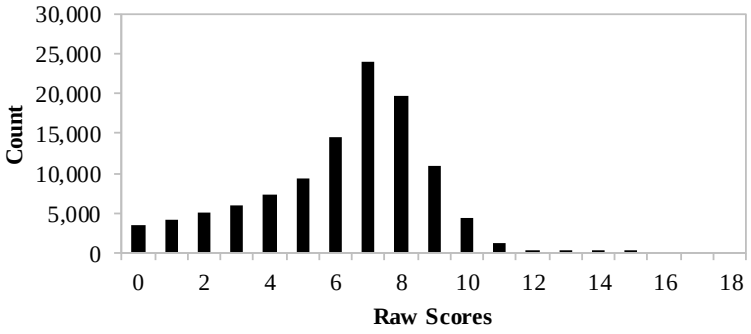
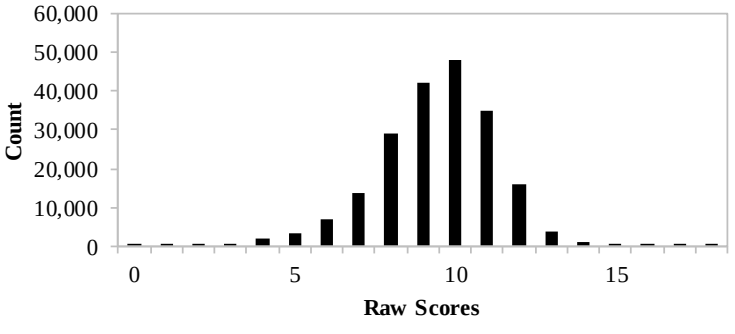


Table 2.3.3.4.2

Raw Score Descriptive Statistics: Writ 6-8 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	82,824	0	16	8.90	1.89
7	65,451	0	18	9.50	1.80
8	53,698	0	18	9.97	1.79
Total	201,973	0	18	9.38	1.88

Figure 2.3.3.4.2
Raw Scores: Writ 6-8 B/C S501 Online



2.3.3.5 *Grades 9–12*

Table 2.3.3.5.1

Raw Score Descriptive Statistics: Writ 9-12 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	47,399	0	16	6.42	3.24
10	29,664	0	16	7.36	2.89
11	21,105	0	17	7.92	2.73
12	16,000	0	16	8.07	2.79
Total	114,168	0	17	7.17	3.07

Figure 2.3.3.5.1
Raw Scores: Writ 9-12 A S501 Online

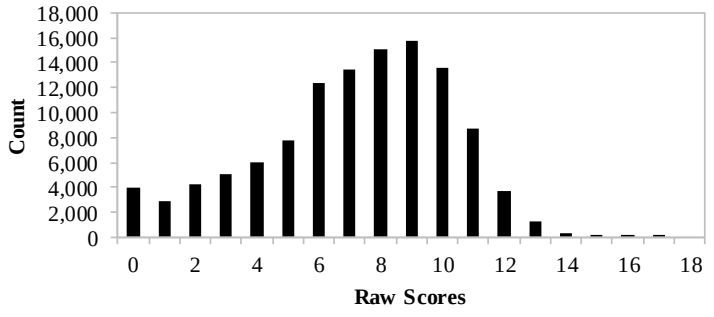
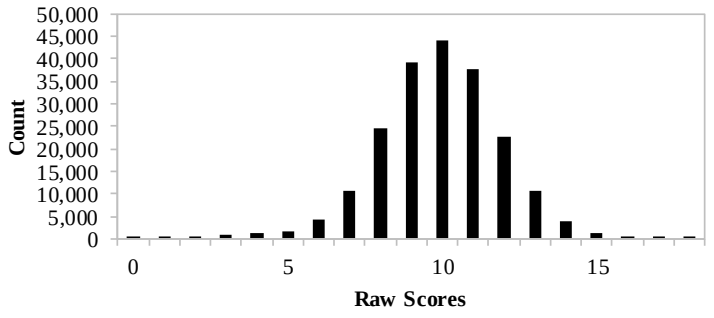


Table 2.3.3.5.2

Raw Score Descriptive Statistics: Writ 9-12 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	58,121	0	17	9.65	2.01
10	53,778	0	18	9.82	2.01
11	48,175	0	18	10.00	2.01
12	43,699	0	17	9.99	2.06
Total	203,773	0	18	9.85	2.03

Figure 2.3.3.5.2
Raw Scores: Writ 9-12 B/C S501 Online



2.3.4 **Speaking**

2.3.4.1 *Grade 1*

Table 2.3.4.1.1

Raw Score Descriptive Statistics: Spek 1 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	7,109	0	6	4.52	2.04
Total	7,109	0	6	4.52	2.04

Figure 2.3.4.1.1
Raw Scores: Spek 1 Pre-A S501 Online

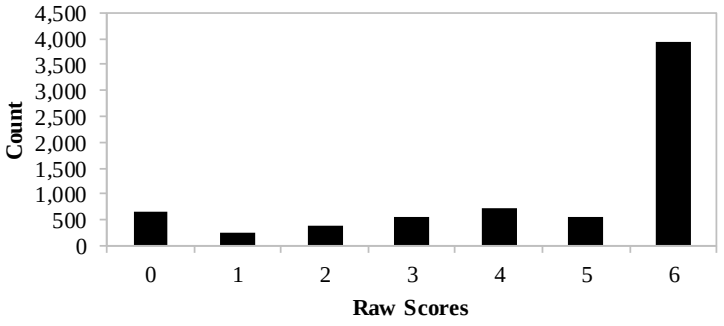


Table 2.3.4.1.2

Raw Score Descriptive Statistics: Spek 1 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	67,864	0	18	10.67	3.22
Total	67,864	0	18	10.67	3.22

Figure 2.3.4.1.2
Raw Scores: Spek 1 A S501 Online

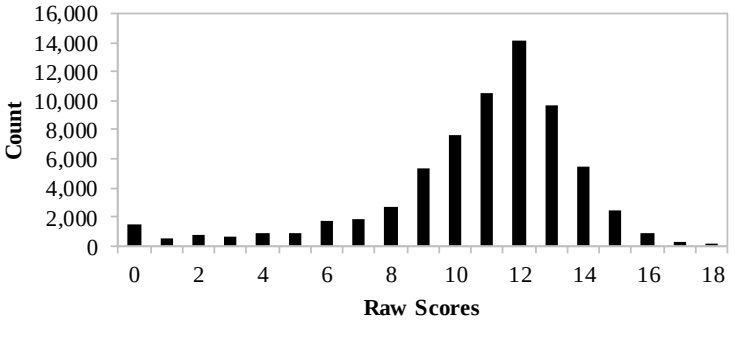
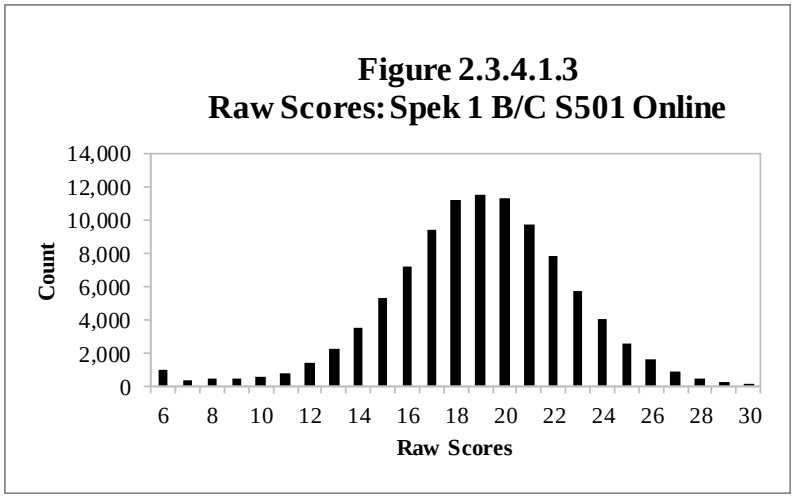


Table 2.3.4.1.3

Raw Score Descriptive Statistics: Spek 1 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	99,910	6	30	18.88	3.81
Total	99,910	6	30	18.88	3.81



2.3.4.2 *Grades 2–3*

Table 2.3.4.2.1

Raw Score Descriptive Statistics: Spek 2-3 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	7,246	0	6	5.08	1.74
3	9,858	0	6	5.07	1.74
Total	17,104	0	6	5.08	1.74

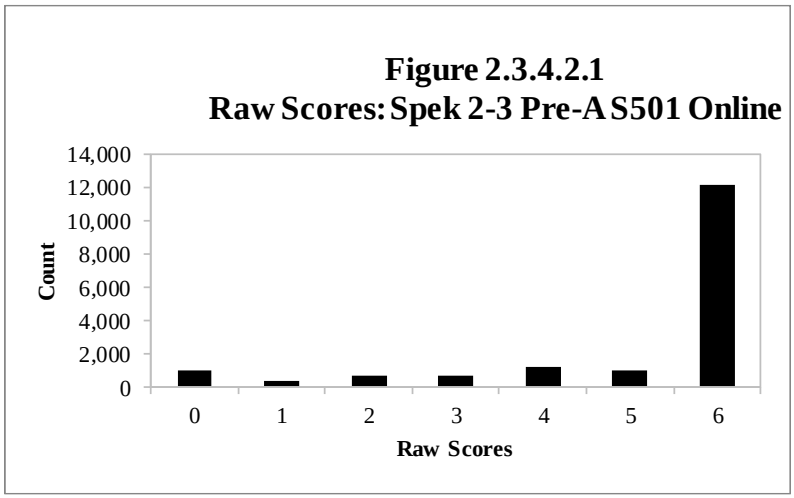


Table 2.3.4.2.2

Raw Score Descriptive Statistics: Spek 2-3 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	43,839	0	18	10.96	2.94
3	38,318	0	18	11.92	2.57
Total	82,157	0	18	11.41	2.81

Figure 2.3.4.2.2
Raw Scores: Spek 2-3 A S501 Online

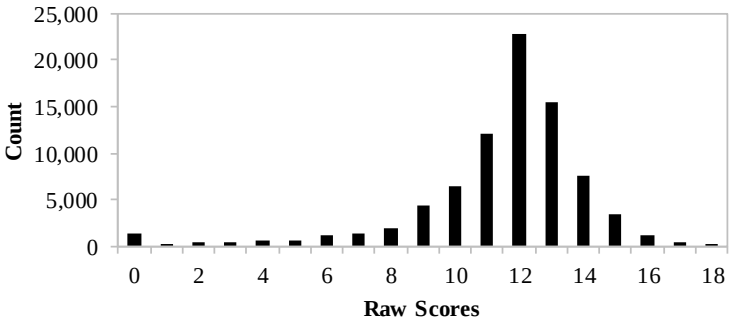
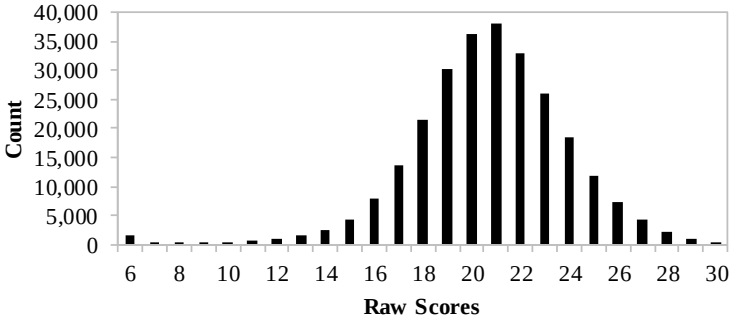


Table 2.3.4.2.3

Raw Score Descriptive Statistics: Spek 2-3 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	131,319	6	30	19.95	3.28
3	133,504	6	30	21.35	3.14
Total	264,823	6	30	20.66	3.29

Figure 2.3.4.2.3
Raw Scores: Spek 2-3 B/C S501 Online



2.3.4.3 *Grades 4–5*

Table 2.3.4.3.1

Raw Score Descriptive Statistics: Spek 4-5 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	2,303	0	6	4.42	2.09
5	4,067	0	6	4.65	1.99
Total	6,370	0	6	4.57	2.03

Figure 2.3.4.3.1
Raw Scores: Spek 4-5 Pre-A S501 Online

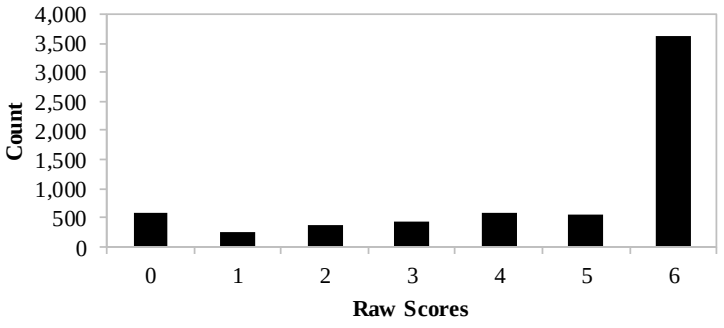


Table 2.3.4.3.2

Raw Score Descriptive Statistics: Spek 4-5 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	17,785	0	18	9.89	3.10
5	13,884	0	18	10.27	2.96
Total	31,669	0	18	10.06	3.04

Figure 2.3.4.3.2
Raw Scores: Spek 4-5 A S501 Online

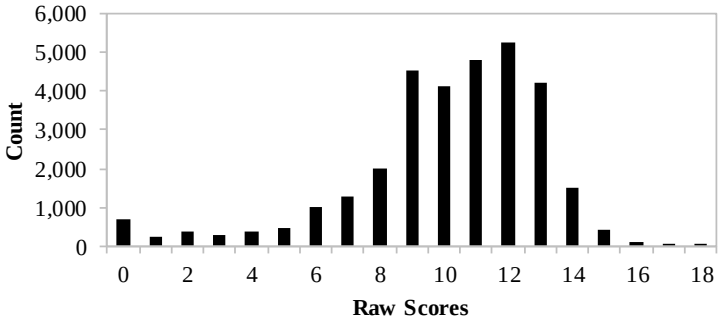
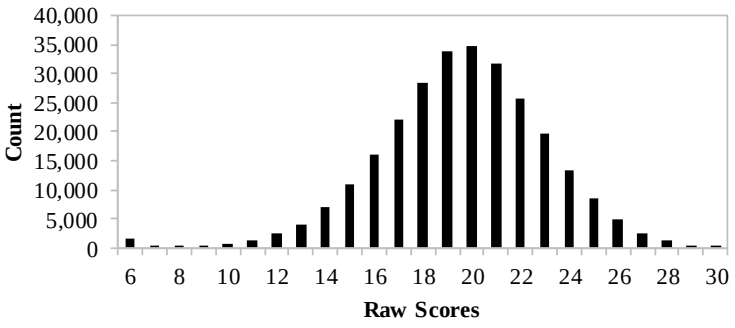


Table 2.3.4.3.3

Raw Score Descriptive Statistics: Spek 4-5 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	151,663	6	30	19.47	3.45
5	121,089	6	30	19.68	3.50
Total	272,752	6	30	19.56	3.47

Figure 2.3.4.3.3
Raw Scores: Spek 4-5 B/C S501 Online



2.3.4.4 *Grades 6–8*

Table 2.3.4.4.1

Raw Score Descriptive Statistics: Spek 6-8 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	2,259	0	6	5.03	1.80
7	3,570	0	6	5.01	1.83
8	3,704	0	6	5.04	1.79
Total	9,533	0	6	5.03	1.81

Figure 2.3.4.4.1
Raw Scores: Spek 6-8 Pre-A S501 Online

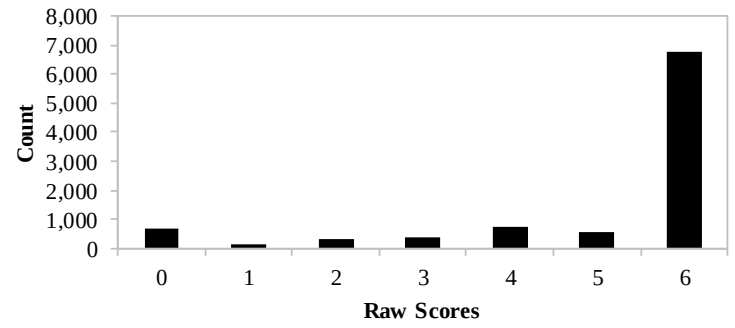


Table 2.3.4.4.2

Raw Score Descriptive Statistics: Spek 6-8 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	19,046	0	17	10.11	2.99
7	16,092	0	17	9.95	3.06
8	27,087	0	18	10.70	2.99
Total	62,225	0	18	10.33	3.03

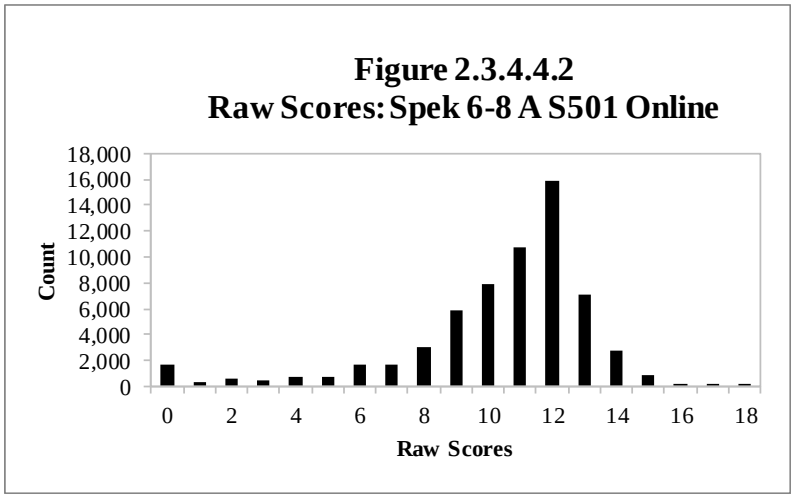
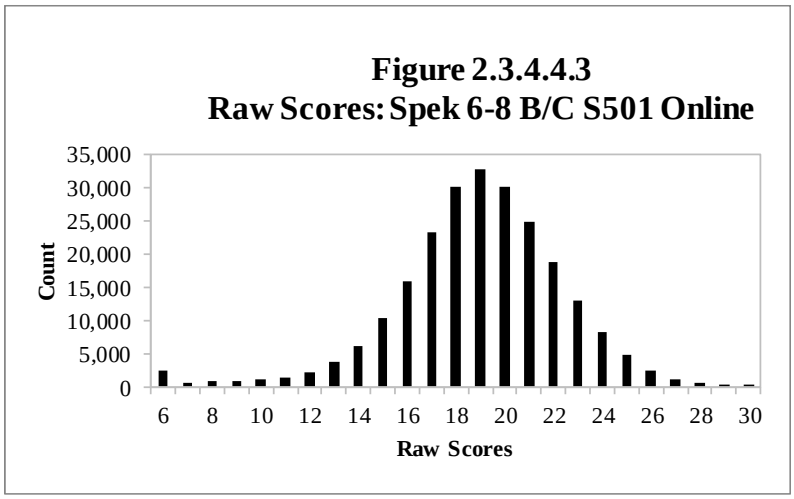


Table 2.3.4.4.3

Raw Score Descriptive Statistics: Spek 6-8 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	93,000	6	30	18.56	3.28
7	83,301	6	30	18.78	3.52
8	59,520	6	30	19.68	3.54
Total	235,821	6	30	18.92	3.46



2.3.4.5 *Grades 9–12*

Table 2.3.4.5.1

Raw Score Descriptive Statistics: Spek 9-12 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	6,238	0	6	4.88	1.84
10	5,280	0	6	5.23	1.64
11	4,105	0	6	5.33	1.61
12	4,266	0	6	5.29	1.72
Total	19,889	0	6	5.15	1.72

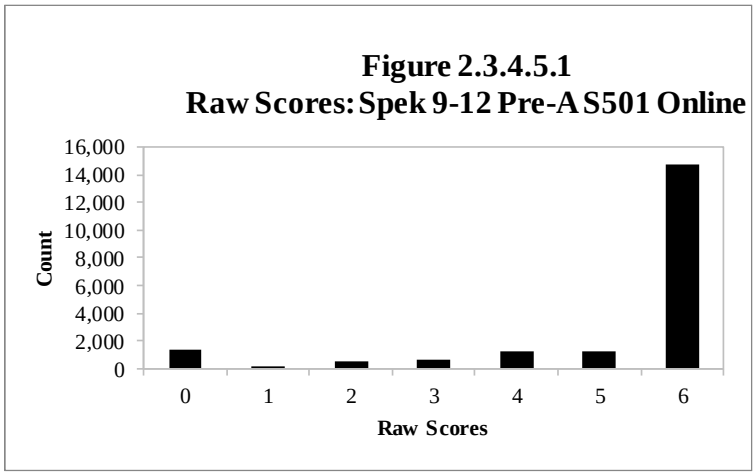


Table 2.3.4.5.2

Raw Score Descriptive Statistics: Spek 9-12 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	57,891	0	18	10.39	3.22
10	32,475	0	18	10.74	3.09
11	13,504	0	17	10.40	3.24
12	25,076	0	18	11.45	3.22
Total	128,946	0	18	10.68	3.21

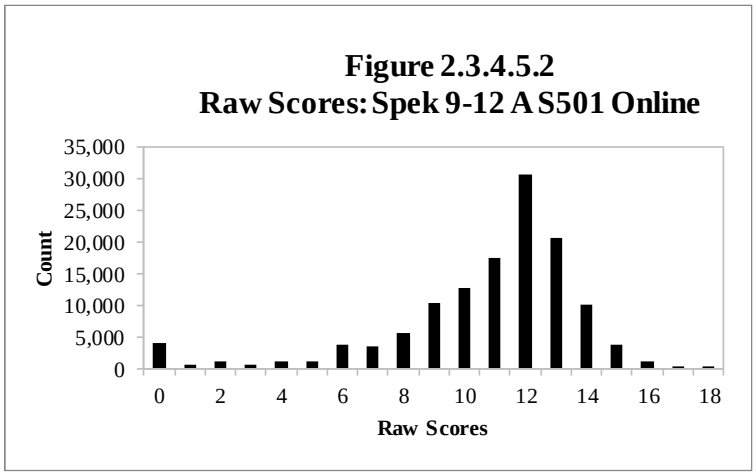
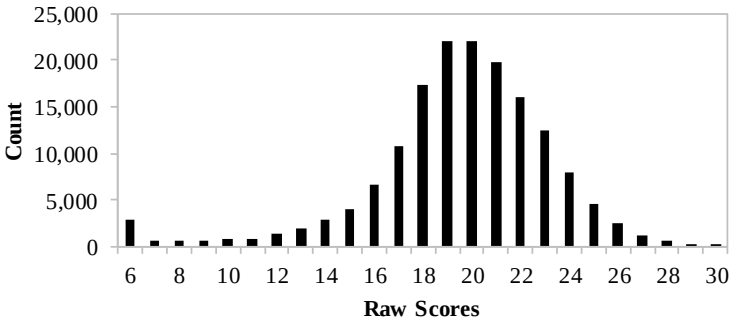


Table 2.3.4.5.3

Raw Score Descriptive Statistics: Spek 9-12 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	38,864	6	30	19.48	3.51
10	43,380	6	30	19.51	3.67
11	49,555	6	30	19.25	3.90
12	28,918	6	30	19.92	3.90
Total	160,717	6	30	19.50	3.76

Figure 2.3.4.5.3
Raw Scores: Spek 9-12 B/C S501 Online



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**

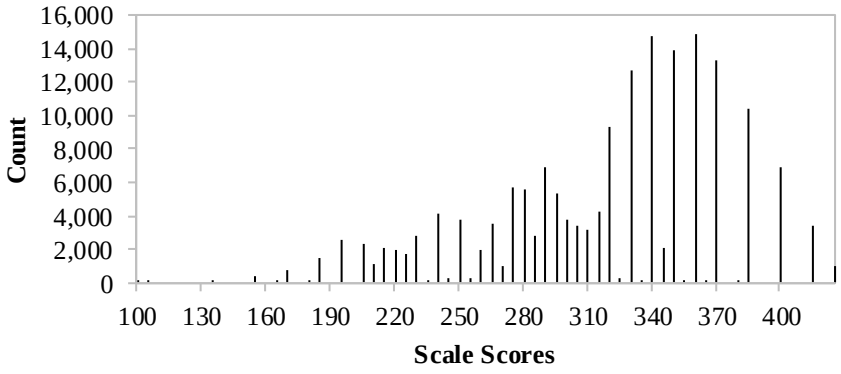
2.4.1.1 *Grade 1*

Table 2.4.1.1

Scale Score Descriptive Statistics: List 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	176,572	104	429	320.27	55.12
Total	176,572	104	429	320.27	55.12

Figure 2.4.1.1
Scale Scores: List 1 S501 Online

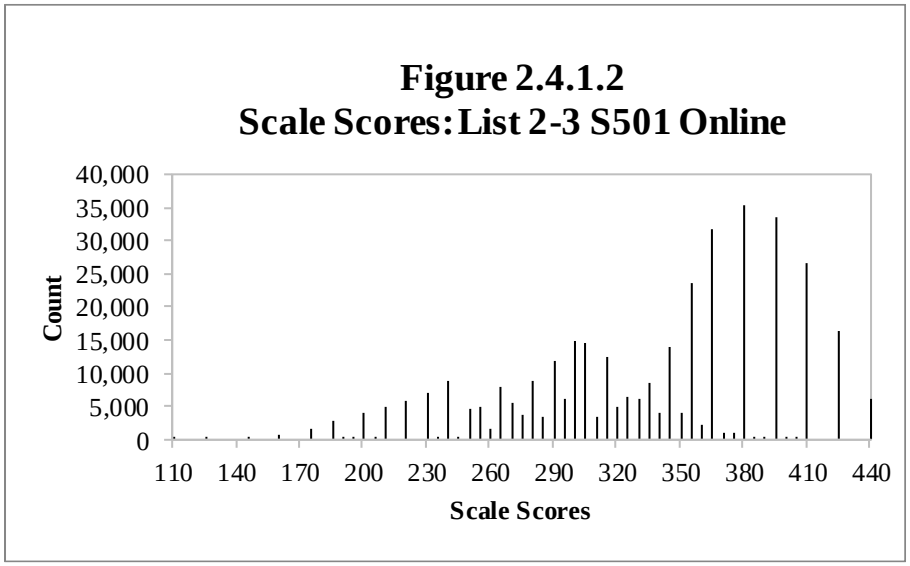


2.4.1.2 *Grades 2–3*

Table 2.4.1.2

Scale Score Descriptive Statistics: List 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	183,889	112	442	325.43	59.94
3	182,714	112	442	350.51	60.36
Total	366,603	112	442	337.93	61.44

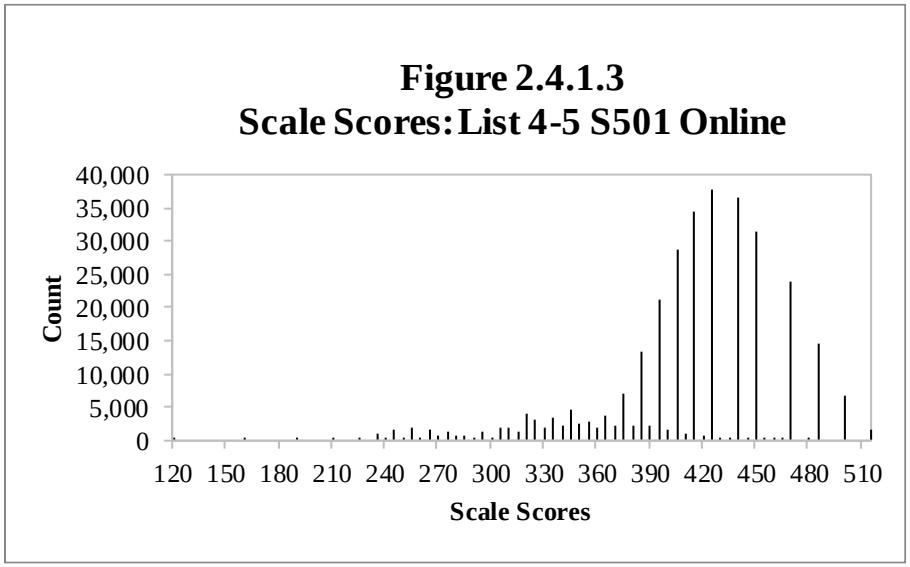


2.4.1.3 *Grades 4–5*

Table 2.4.1.3

Scale Score Descriptive Statistics: List 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	174,730	120	518	410.91	49.73
5	140,985	120	518	418.14	53.77
Total	315,715	120	518	414.14	51.70



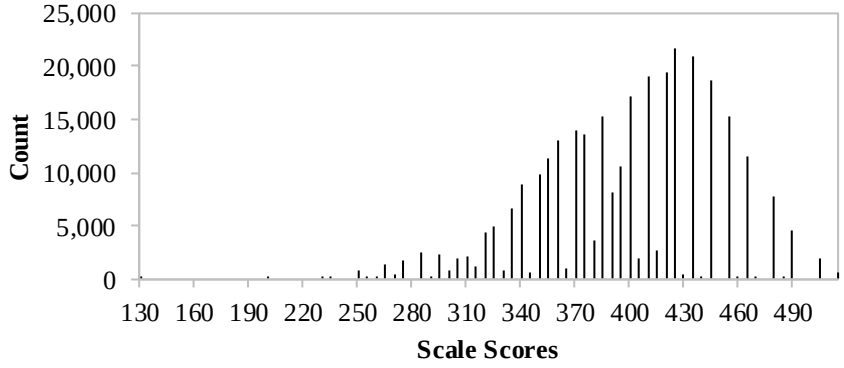
2.4.1.4 *Grades 6–8*

Table 2.4.1.4

Scale Score Descriptive Statistics: List 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	114,021	132	519	396.10	43.60
7	102,671	132	519	399.19	48.85
8	89,927	132	519	404.76	52.23
Total	306,619	132	519	399.67	48.15

Figure 2.4.1.4
Scale Scores: List 6-8 S501 Online



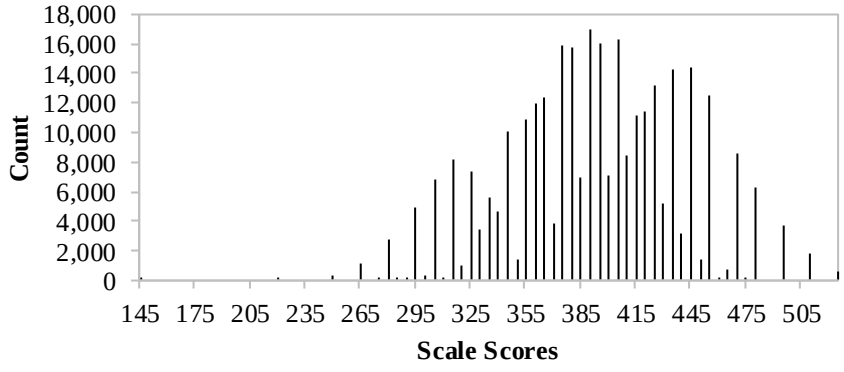
2.4.1.5 *Grades 9–12*

Table 2.4.1.5

Scale Score Descriptive Statistics: List 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	102,240	148	526	386.04	48.04
10	81,296	220	526	394.84	48.29
11	67,599	148	526	400.77	47.12
12	58,410	148	526	402.52	47.06
Total	309,545	148	526	394.68	48.18

Figure 2.4.1.5
Scale Scores: List 9-12 S501 Online



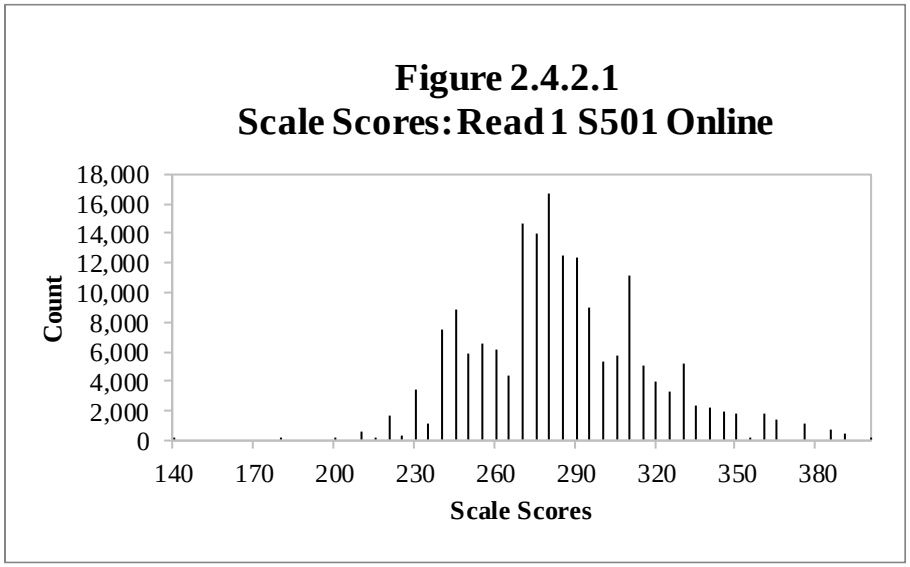
2.4.2 **Reading**

2.4.2.1 *Grade 1*

Table 2.4.2.1

Scale Score Descriptive Statistics: Read 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	179,739	141	403	287.07	32.24
Total	179,739	141	403	287.07	32.24

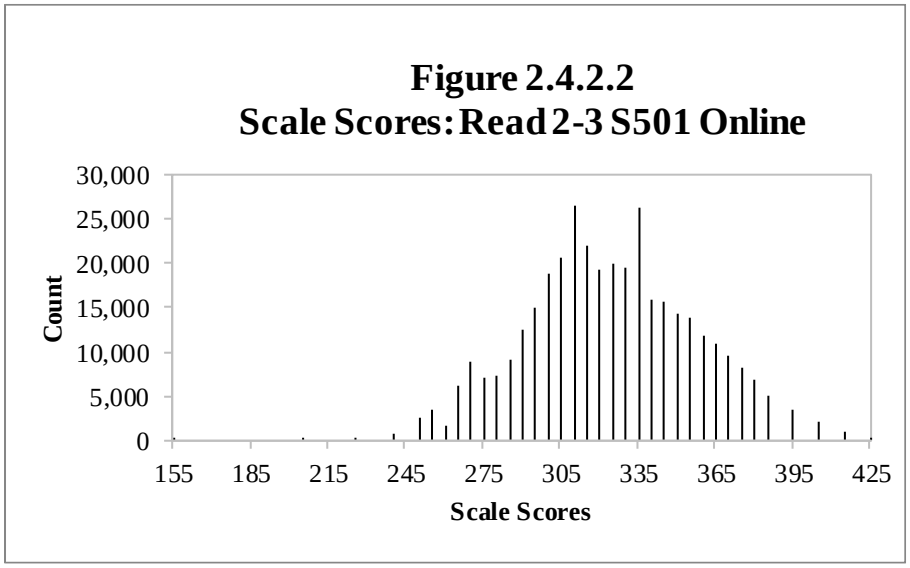


2.4.2.2 *Grades 2–3*

Table 2.4.2.2

Scale Score Descriptive Statistics: Read 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	184,150	158	425	318.68	29.12
3	182,462	158	425	333.06	33.75
Total	366,612	158	425	325.84	32.32

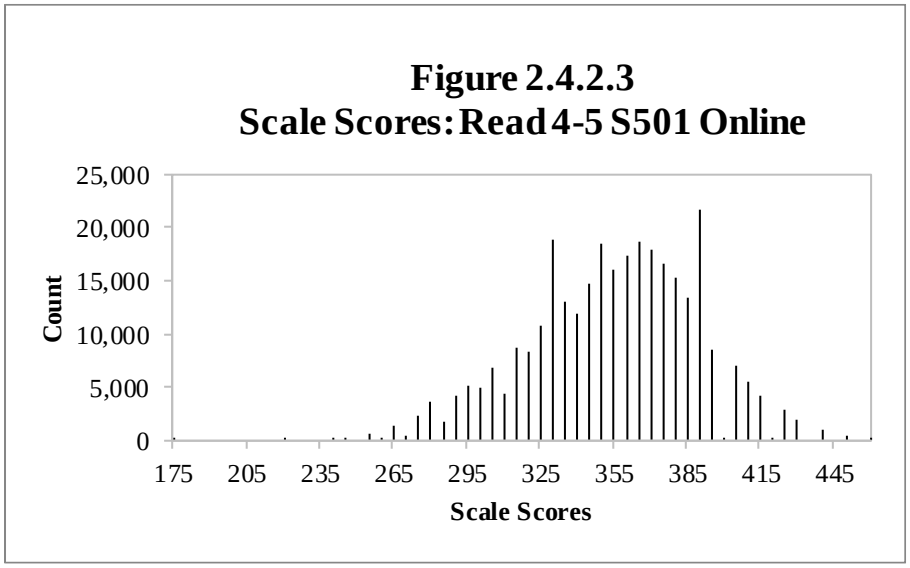


2.4.2.3 *Grades 4–5*

Table 2.4.2.3

Scale Score Descriptive Statistics: Read 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	171,235	175	461	354.68	32.91
5	138,312	175	461	358.34	35.32
Total	309,547	175	461	356.32	34.06



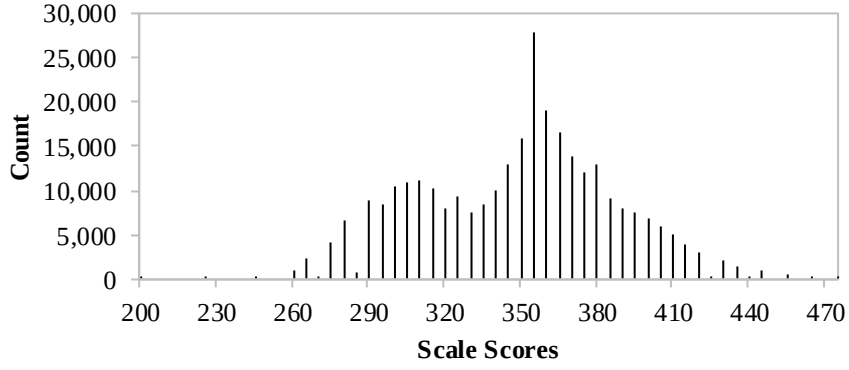
2.4.2.4 *Grades 6–8*

Table 2.4.2.4

Scale Score Descriptive Statistics: Read 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	113,310	200	475	345.28	35.13
7	101,914	200	475	350.92	38.08
8	88,867	200	475	356.83	41.33
Total	304,091	200	475	350.55	38.30

Figure 2.4.2.4
Scale Scores: Read 6-8 S501 Online



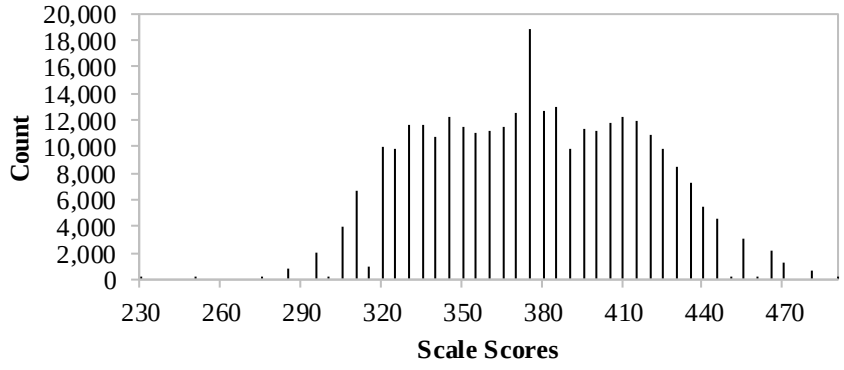
2.4.2.5 *Grades 9–12*

Table 2.4.2.5

Scale Score Descriptive Statistics: Read 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	101,358	233	492	368.81	38.55
10	79,762	233	492	379.28	38.85
11	66,214	254	492	386.23	38.52
12	57,441	233	492	387.97	37.94
Total	304,775	233	492	378.95	39.29

Figure 2.4.2.5
Scale Scores: Read 9-12 S501 Online



2.4.3 **Writing**

2.4.3.1 *Grade 1*

Table 2.4.3.1.1

Scale Score Descriptive Statistics: Writ 1 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	158,459	111	355	247.89	37.79
Total	158,459	111	355	247.89	37.79

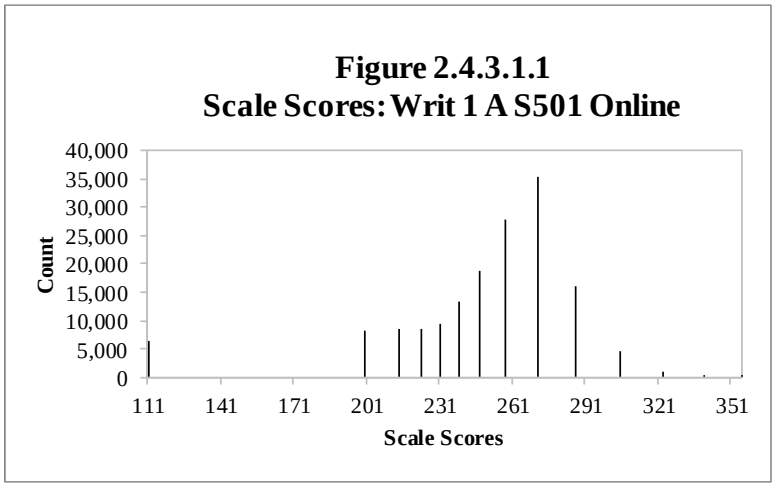


Table 2.4.3.1.2

Scale Score Descriptive Statistics: Writ 1 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	28,391	111	395	295.67	28.38
Total	28,391	111	395	295.67	28.38

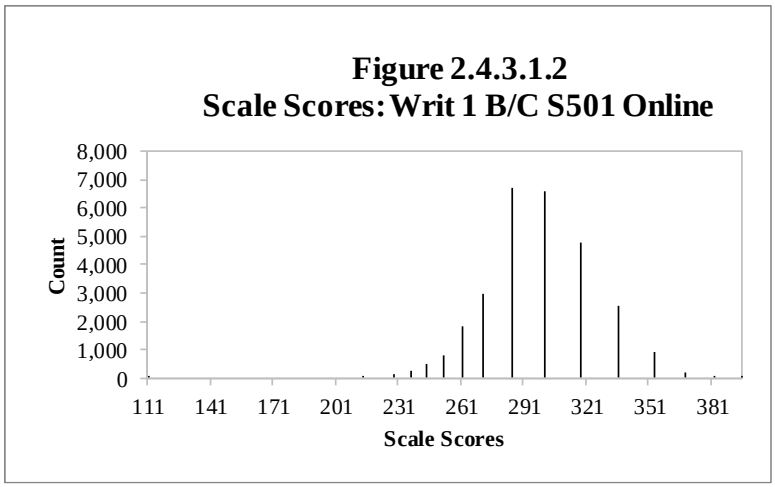
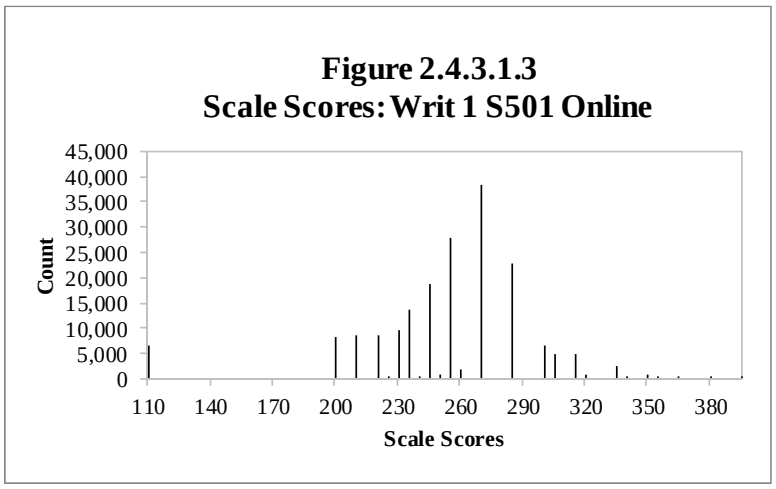


Table 2.4.3.1.3

Scale Score Descriptive Statistics: Writ 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	186,850	111	395	255.15	40.35
Total	186,850	111	395	255.15	40.35



2.4.3.2 *Grades 2–3*

Table 2.4.3.2.1

Scale Score Descriptive Statistics: Writ 2-3 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	56,977	133	399	250.24	42.95
3	38,672	133	385	257.24	44.11
Total	95,649	133	399	253.07	43.56

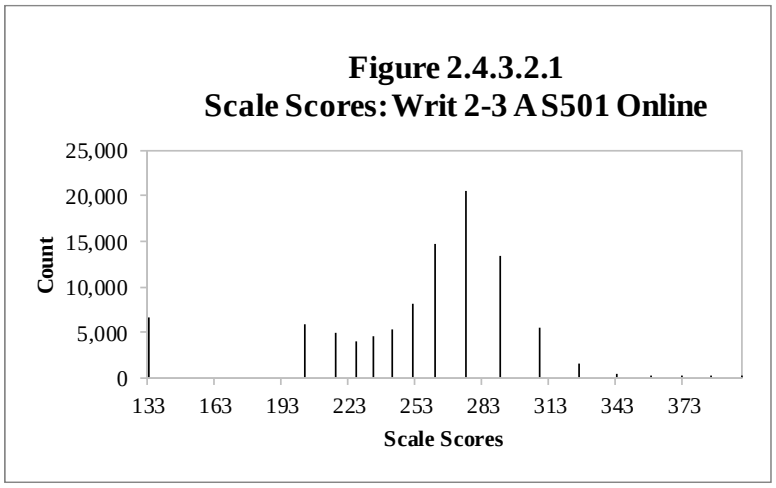


Table 2.4.3.2.2

Scale Score Descriptive Statistics: Writ 2-3 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	137,151	133	414	311.87	28.46
3	153,337	133	467	328.32	25.26
Total	290,488	133	467	320.55	28.05

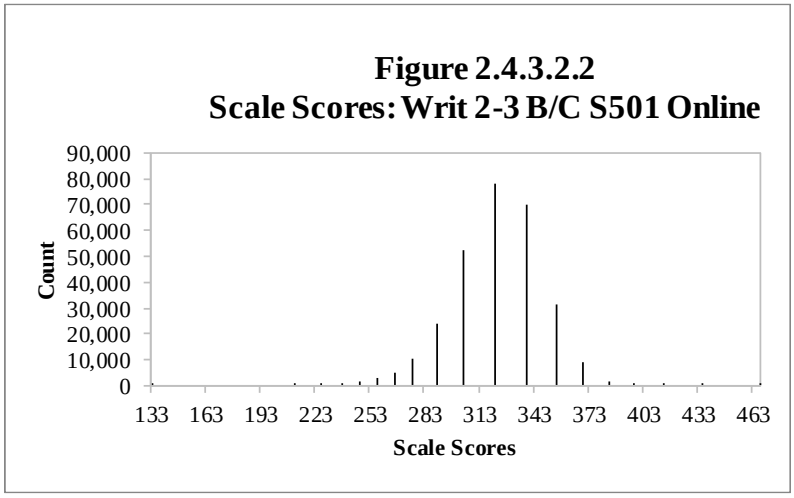
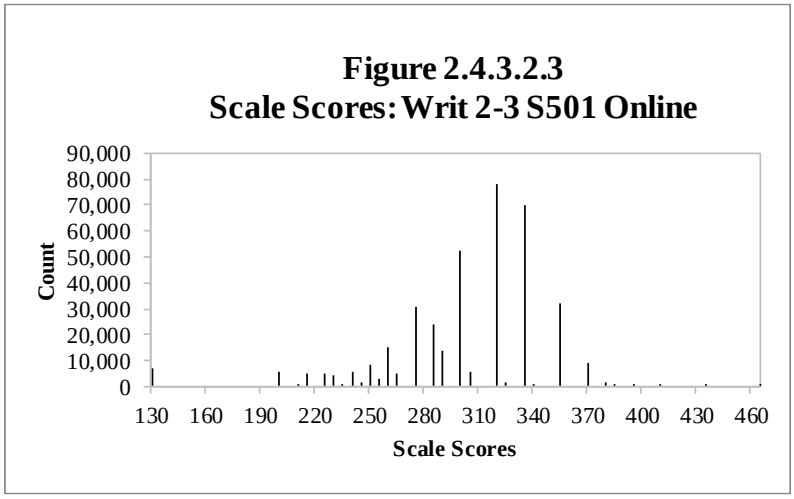


Table 2.4.3.2.3

Scale Score Descriptive Statistics: Writ 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	194,128	133	414	293.78	43.60
3	192,009	133	467	314.00	41.40
Total	386,137	133	467	303.84	43.71



2.4.3.3 *Grades 4–5*

Table 2.4.3.3.1

Scale Score Descriptive Statistics: Writ 4-5 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	24,984	155	405	264.47	52.58
5	24,928	155	418	274.99	49.85
Total	49,912	155	418	269.72	51.50

Figure 2.4.3.3.1
Scale Scores: Writ 4-5 A S501 Online

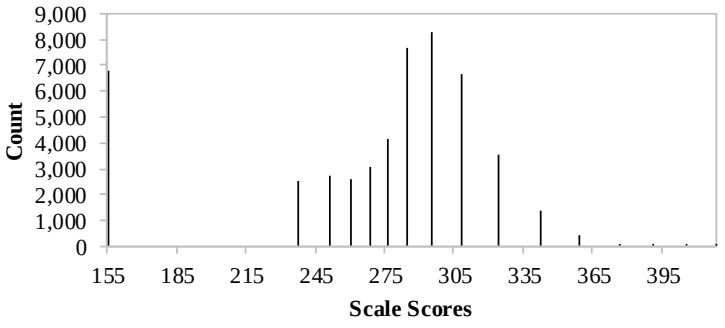


Table 2.4.3.3.2

Scale Score Descriptive Statistics: Writ 4-5 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	150,873	155	487	345.98	30.76
5	117,540	155	487	356.38	29.30
Total	268,413	155	487	350.53	30.57

Figure 2.4.3.3.2
Scale Scores: Writ 4-5 B/C S501 Online

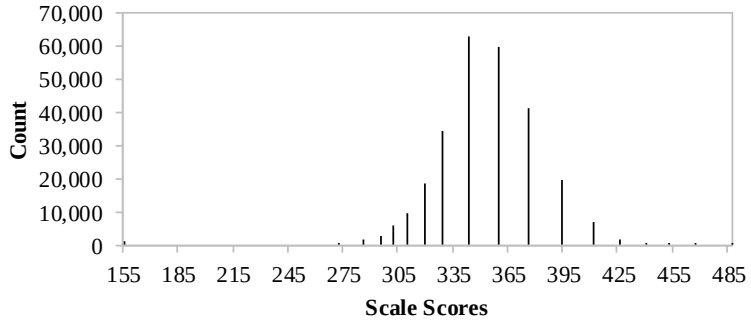
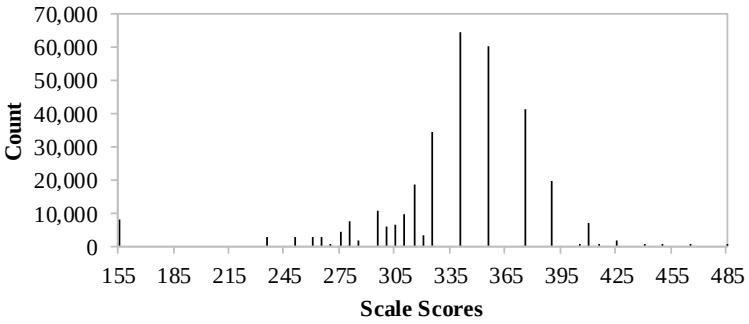


Table 2.4.3.3.3

Scale Score Descriptive Statistics: Writ 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	175,857	155	487	334.40	44.88
5	142,468	155	487	342.14	45.82
Total	318,325	155	487	337.86	45.47

Figure 2.4.3.3.3
Scale Scores: Writ 4-5 S501 Online



2.4.3.4 *Grades 6–8*

Table 2.4.3.4.1

Scale Score Descriptive Statistics: Writ 6-8 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	33,634	188	394	286.46	31.73
7	39,214	188	408	292.59	31.40
8	37,263	188	421	295.79	31.52
Total	110,111	188	421	291.80	31.77

Figure 2.4.3.4.1
Scale Scores: Writ 6-8 A S501 Online

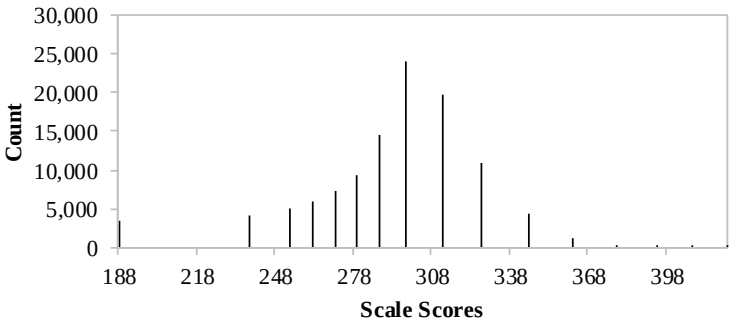


Table 2.4.3.4.2

Scale Score Descriptive Statistics: Writ 6-8 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	82,824	188	441	333.69	27.98
7	65,451	188	492	343.07	27.90
8	53,698	188	492	350.58	28.40
Total	201,973	188	492	341.22	28.90

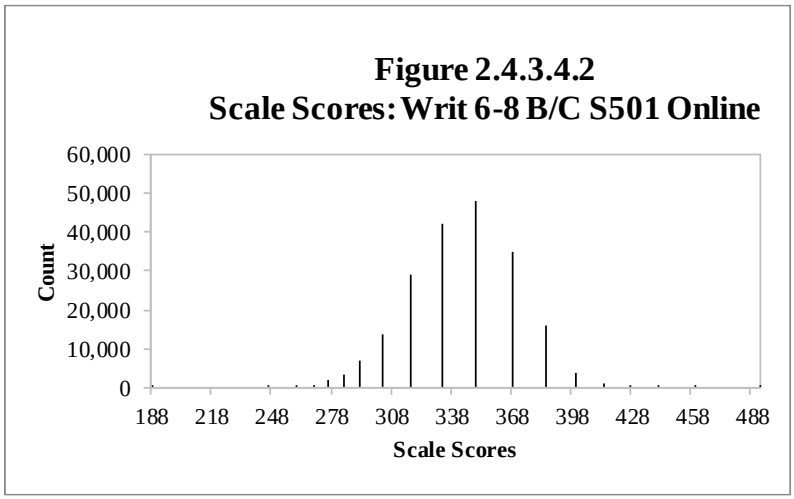
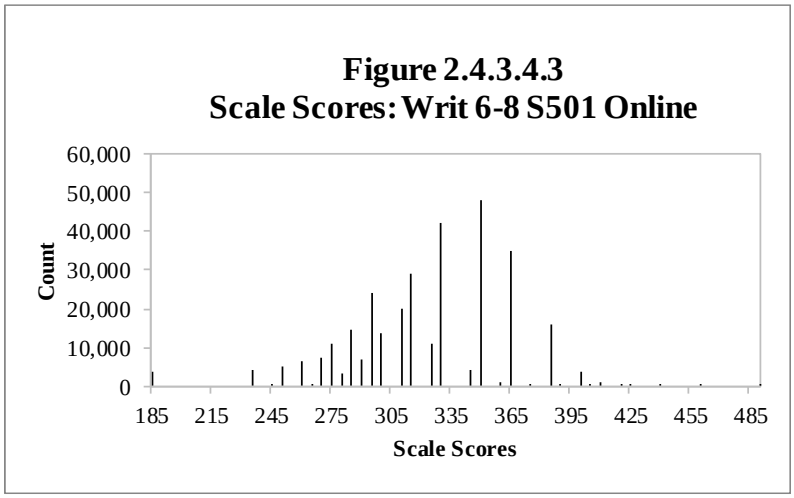


Table 2.4.3.4.3

Scale Score Descriptive Statistics: Writ 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	116,458	188	441	320.05	36.13
7	104,665	188	492	324.15	38.12
8	90,961	188	492	328.13	40.11
Total	312,084	188	492	323.78	38.13



2.4.3.5 *Grades 9–12*

Table 2.4.3.5.1

Scale Score Descriptive Statistics: Writ 9-12 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	47,399	232	452	314.61	40.06
10	29,664	232	452	325.49	37.23
11	21,105	232	472	332.52	36.37
12	16,000	232	452	334.58	37.48
Total	114,168	232	472	323.55	39.16

Figure 2.4.3.5.1
Scale Scores: Writ 9-12 A S501 Online

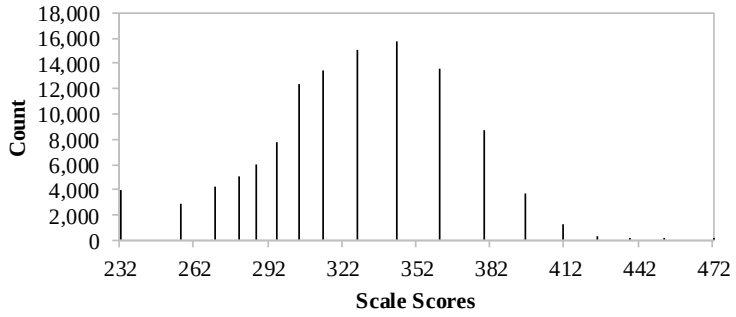


Table 2.4.3.5.2

Scale Score Descriptive Statistics: Writ 9-12 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	58,121	232	476	360.67	30.91
10	53,778	232	508	363.30	31.13
11	48,175	232	508	366.26	31.26
12	43,699	232	476	366.19	31.76
Total	203,773	232	508	363.87	31.32

Figure 2.4.3.5.2
Scale Scores: Writ 9-12 B/C S501 Online

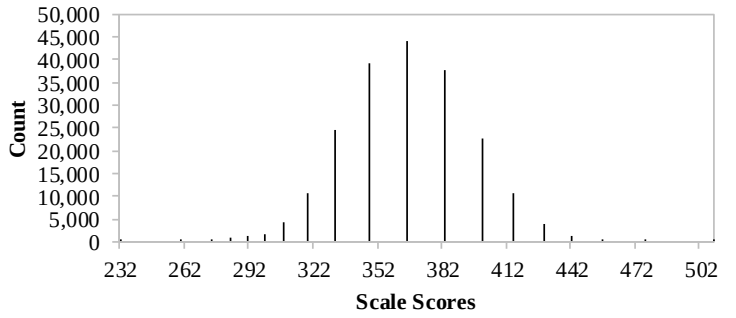
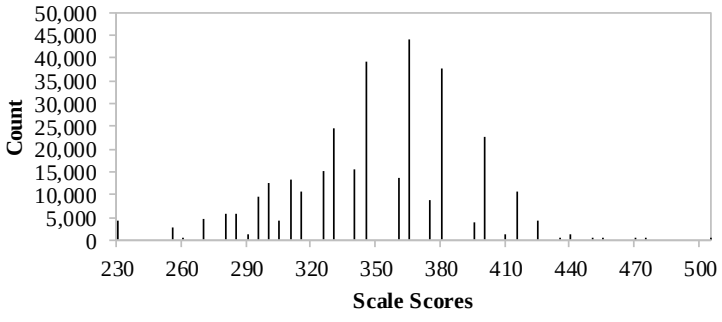


Table 2.4.3.5.3

Scale Score Descriptive Statistics: Writ 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	105,520	232	476	339.98	42.09
10	83,442	232	508	349.85	38.01
11	69,280	232	508	355.98	36.38
12	59,699	232	476	357.72	36.21
Total	317,941	232	508	349.39	39.42

Figure 2.4.3.5.3
Scale Scores: Writ 9-12 S501 Online



2.4.4 **Speaking**

2.4.4.1 *Grade 1*

Table 2.4.4.1.1

Scale Score Descriptive Statistics: Spek 1 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	7,109	106	172	153.72	24.32
Total	7,109	106	172	153.72	24.32

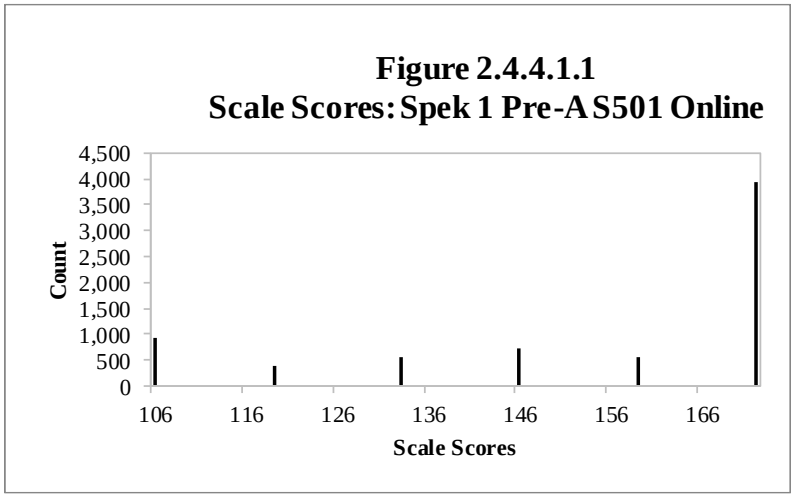


Table 2.4.4.1.2

Scale Score Descriptive Statistics: Spek 1 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	67,864	106	387	238.05	53.81
Total	67,864	106	387	238.05	53.81

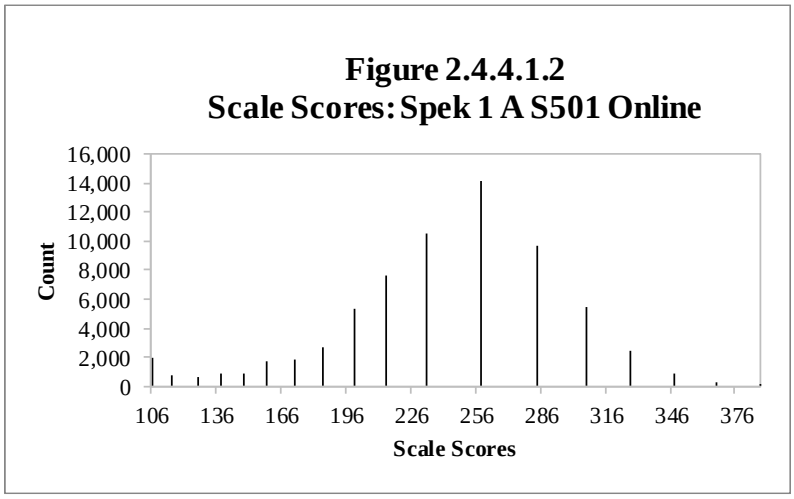


Table 2.4.4.1.3

Scale Score Descriptive Statistics: Spek 1 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	99,910	106	404	273.75	43.65
Total	99,910	106	404	273.75	43.65

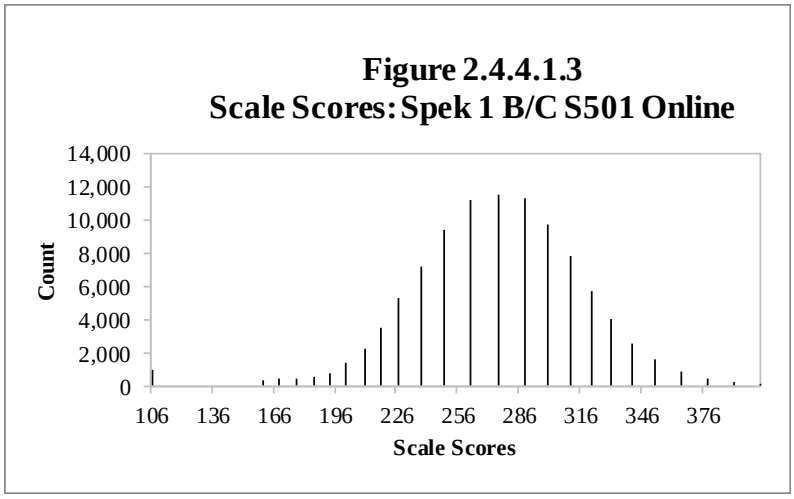
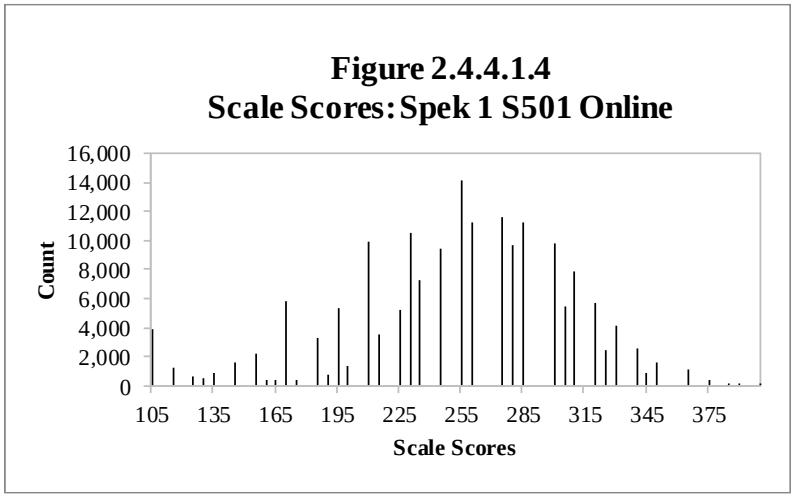


Table 2.4.4.1.4

Scale Score Descriptive Statistics: Spek 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	174,883	106	404	255.02	54.46
Total	174,883	106	404	255.02	54.46



2.4.4.2 *Grades 2–3*

Table 2.4.4.2.1

Scale Score Descriptive Statistics: Spek 2-3 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	7,246	118	166	156.29	17.10
3	9,858	118	166	156.09	17.15
Total	17,104	118	166	156.17	17.13

Figure 2.4.4.2.1
Scale Scores: Spek 2-3 Pre-A S501 Online

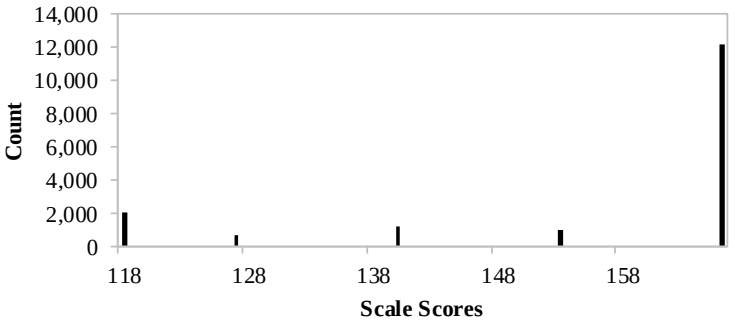


Table 2.4.4.2.2

Scale Score Descriptive Statistics: Spek 2-3 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	43,839	118	387	240.09	48.44
3	38,318	118	387	258.68	46.13
Total	82,157	118	387	248.76	48.28

Figure 2.4.4.2.2
Scale Scores: Spek 2-3 A S501 Online

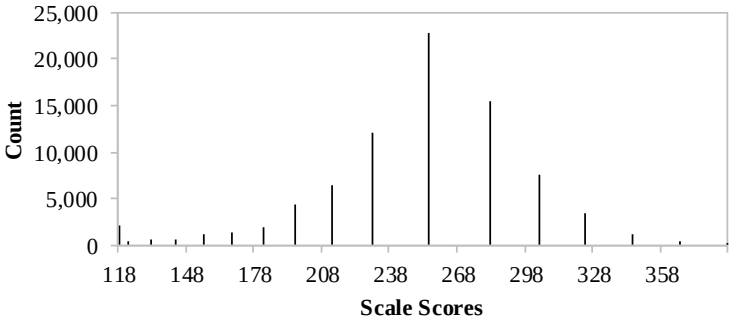


Table 2.4.4.2.3

Scale Score Descriptive Statistics: Spek 2-3 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	131,319	118	425	283.21	37.62
3	133,504	118	425	299.34	36.08
Total	264,823	118	425	291.35	37.72

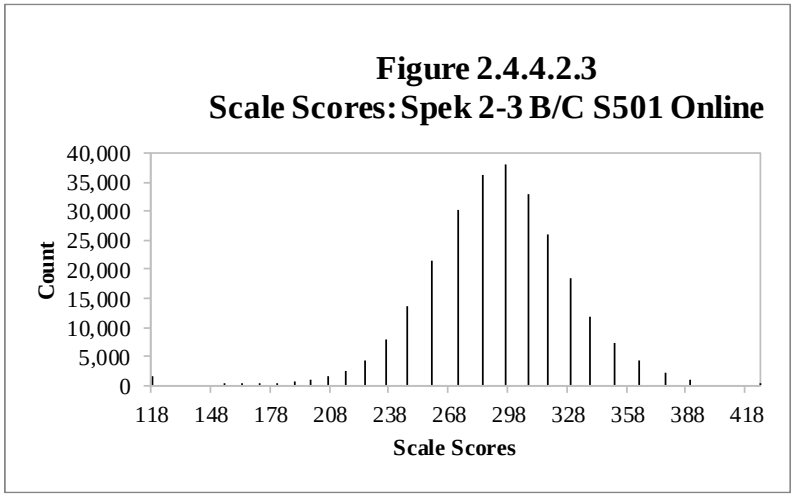
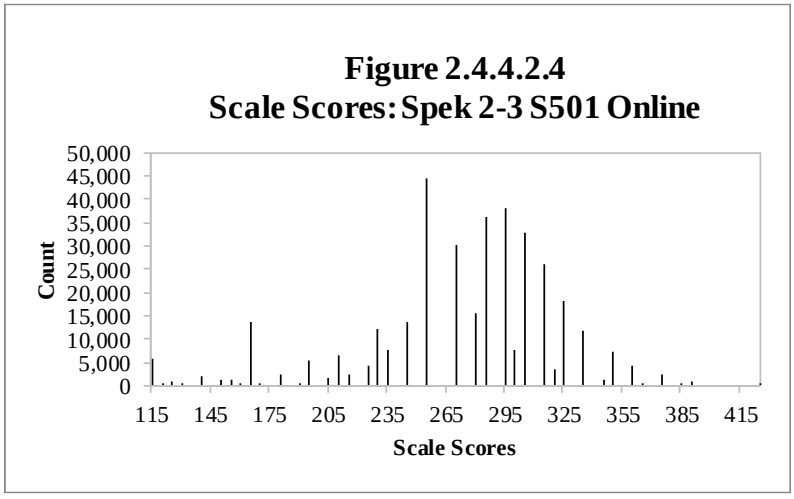


Table 2.4.4.2.4

Scale Score Descriptive Statistics: Spek 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	182,404	118	425	267.81	49.43
3	181,680	118	425	282.99	51.15
Total	364,084	118	425	275.39	50.87



2.4.4.3 *Grades 4–5*

Table 2.4.4.3.1

Scale Score Descriptive Statistics: Spek 4-5 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	2,303	130	189	169.84	23.76
5	4,067	130	189	172.56	22.68
Total	6,370	130	189	171.58	23.11

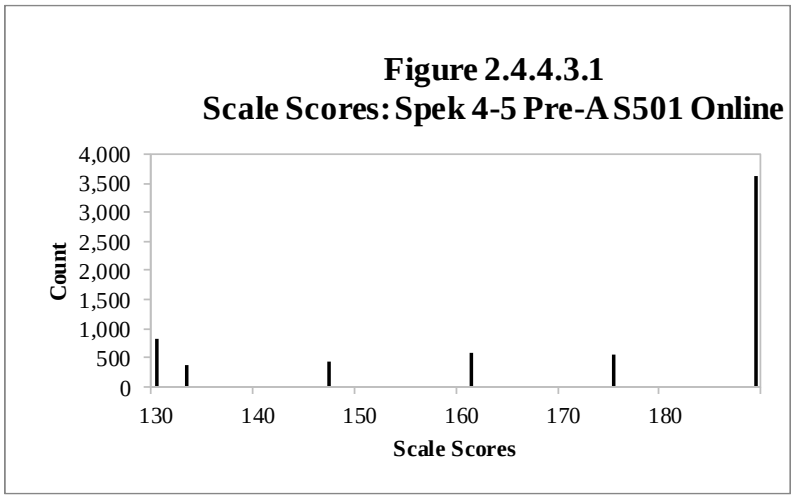


Table 2.4.4.3.2

Scale Score Descriptive Statistics: Spek 4-5 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	17,785	130	429	254.02	53.13
5	13,884	130	429	260.98	52.09
Total	31,669	130	429	257.07	52.79

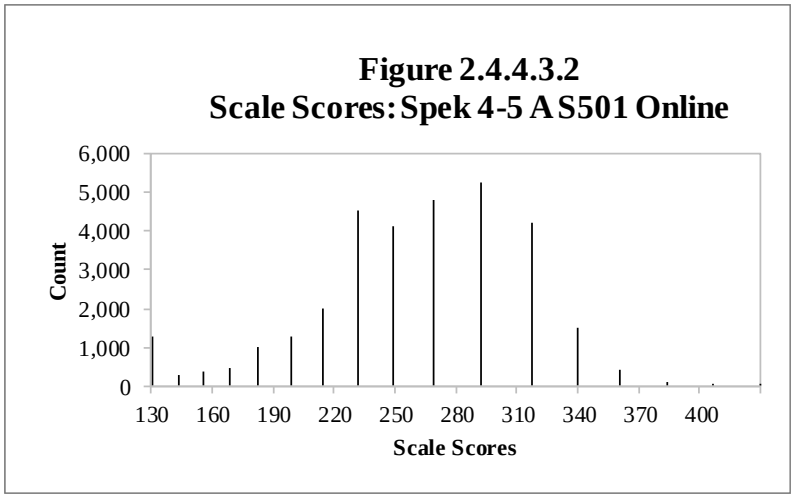


Table 2.4.4.3.3

Scale Score Descriptive Statistics: Spek 4-5 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	151,663	130	448	322.43	40.33
5	121,089	130	448	324.80	41.01
Total	272,752	130	448	323.48	40.65

Figure 2.4.4.3.3
Scale Scores: Spek 4-5 B/C S501 Online

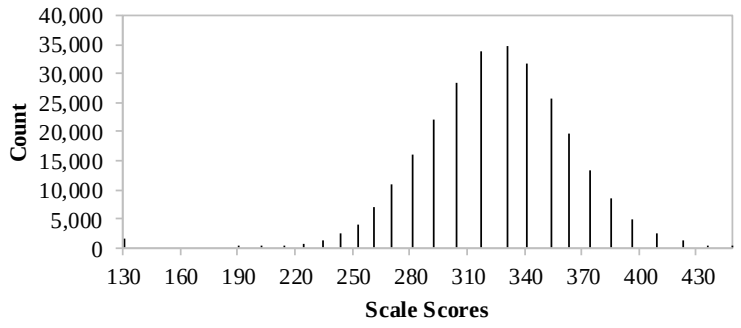
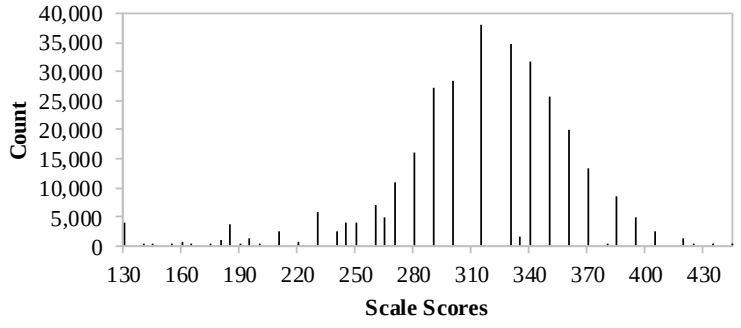


Table 2.4.4.3.4

Scale Score Descriptive Statistics: Spek 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	171,751	130	448	313.30	49.49
5	139,040	130	448	313.98	52.13
Total	310,791	130	448	313.60	50.69

Figure 2.4.4.3.4
Scale Scores: Spek 4-5 S501 Online



2.4.4.4 *Grades 6–8*

Table 2.4.4.4.1

Scale Score Descriptive Statistics: Spek 6-8 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	2,259	148	210	198.12	20.99
7	3,570	148	210	197.88	21.14
8	3,704	148	210	198.29	20.78
Total	9,533	148	210	198.10	20.97

Figure 2.4.4.4.1
Scale Scores: Spek 6-8 Pre-A S501 Online

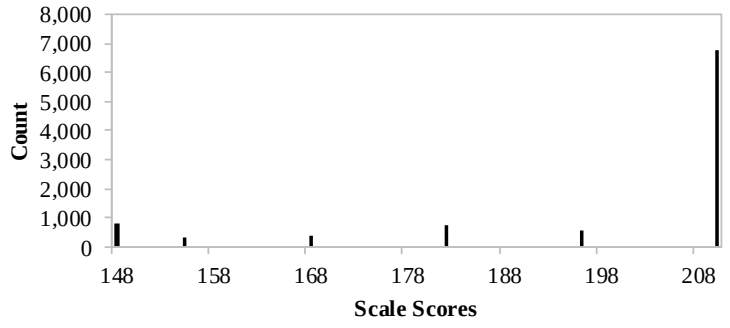


Table 2.4.4.4.2

Scale Score Descriptive Statistics: Spek 6-8 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	19,046	148	416	273.84	49.25
7	16,092	148	416	270.84	49.53
8	27,087	148	436	285.73	50.70
Total	62,225	148	436	278.24	50.40

Figure 2.4.4.4.2
Scale Scores: Spek 6-8 A S501 Online

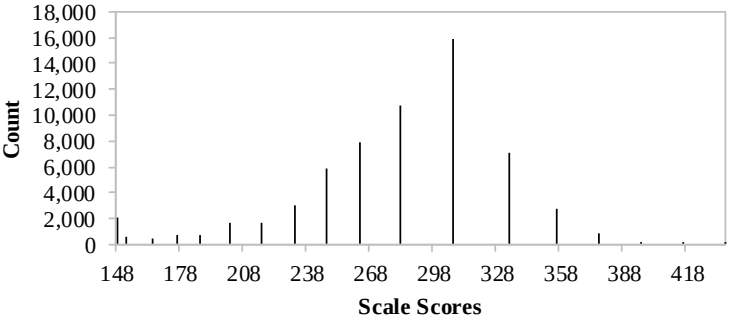


Table 2.4.4.4.3

Scale Score Descriptive Statistics: Spek 6-8 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	93,000	148	463	325.46	39.08
7	83,301	148	463	328.01	41.82
8	59,520	148	463	338.36	42.10
Total	235,821	148	463	329.62	41.16

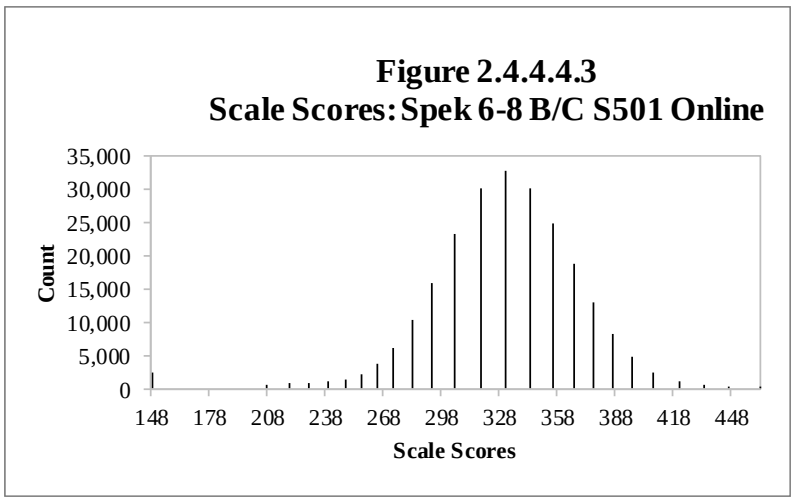
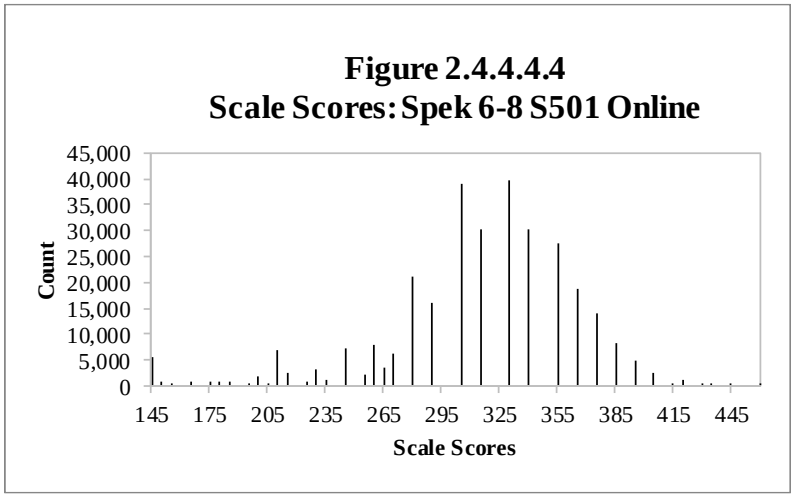


Table 2.4.4.4.4

Scale Score Descriptive Statistics: Spek 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	114,305	148	463	314.34	47.92
7	102,963	148	463	314.56	52.25
8	90,311	148	463	316.83	55.93
Total	307,579	148	463	315.15	51.84



2.4.4.5 *Grades 9–12*

Table 2.4.4.5.1

Scale Score Descriptive Statistics: Spek 9-12 Pre-A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	6,238	172	222	209.39	18.78
10	5,280	172	222	213.53	16.48
11	4,105	172	222	214.88	15.57
12	4,266	172	222	214.80	16.13
Total	19,889	172	222	212.78	17.15

Figure 2.4.4.5.1
Scale Scores: Spek 9-12 Pre-AS501 Online

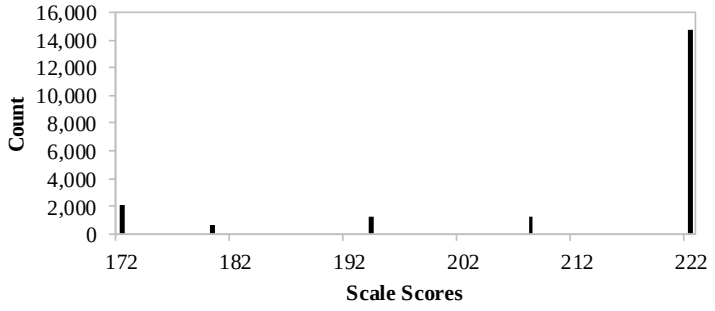


Table 2.4.4.5.2

Scale Score Descriptive Statistics: Spek 9-12 A S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	57,891	172	442	285.00	51.59
10	32,475	172	442	291.15	49.92
11	13,504	172	421	285.18	49.91
12	25,076	172	442	306.80	52.21
Total	128,946	172	442	290.81	51.79

Figure 2.4.4.5.2
Scale Scores: Spek 9-12 A S501 Online

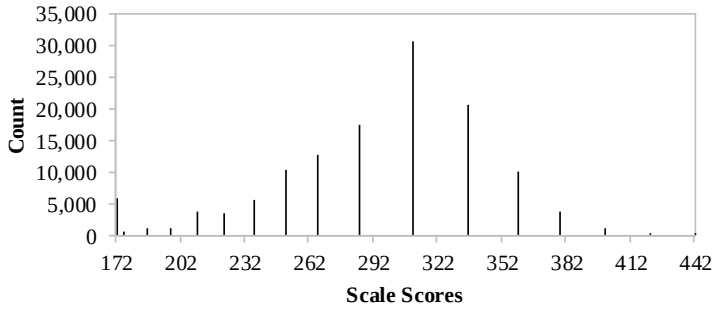


Table 2.4.4.5.3

Scale Score Descriptive Statistics: Spek 9-12 B/C S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	38,864	172	476	339.76	40.68
10	43,380	172	476	340.01	42.48
11	49,555	172	476	337.03	45.07
12	28,918	172	476	344.88	45.07
Total	160,717	172	476	339.91	43.43

Figure 2.4.4.5.3
Scale Scores: Spek 9-12 B/C S501 Online

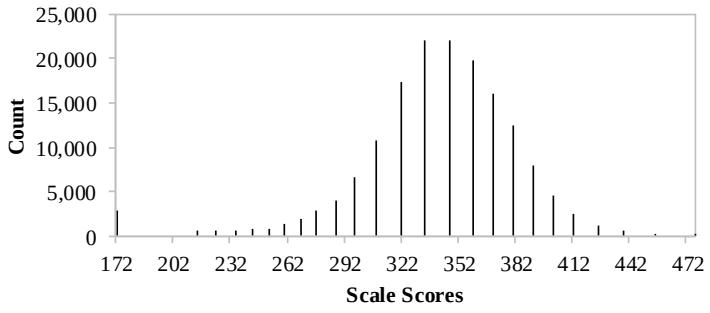
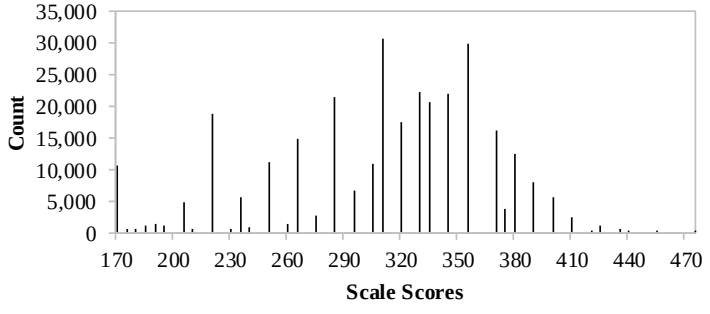


Table 2.4.4.5.4

Scale Score Descriptive Statistics: Spek 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	102,993	172	476	301.08	57.97
10	81,135	172	476	312.22	56.61
11	67,164	172	476	319.14	56.10
12	58,260	172	476	318.96	58.24
Total	309,552	172	476	311.29	57.78

Figure 2.4.4.5.4
Scale Scores: Spek 9-12 S501 Online



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

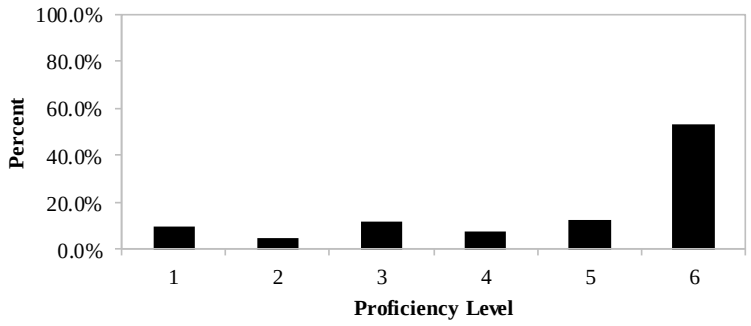
2.5.1.1 *Grade 1*

Table 2.5.1.1

Proficiency Level Distribution: List 1 S501 Online

	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	17,640	9.99%	17,640	9.99%
2	8,433	4.78%	8,433	4.78%
3	21,438	12.14%	21,438	12.14%
4	13,344	7.56%	13,344	7.56%
5	22,142	12.54%	22,142	12.54%
6	93,575	53.00%	93,575	53.00%
Total	176,572	100.00%	176,572	100.00%

Figure 2.5.1.1
Proficiency Level: List 1 S501 Online



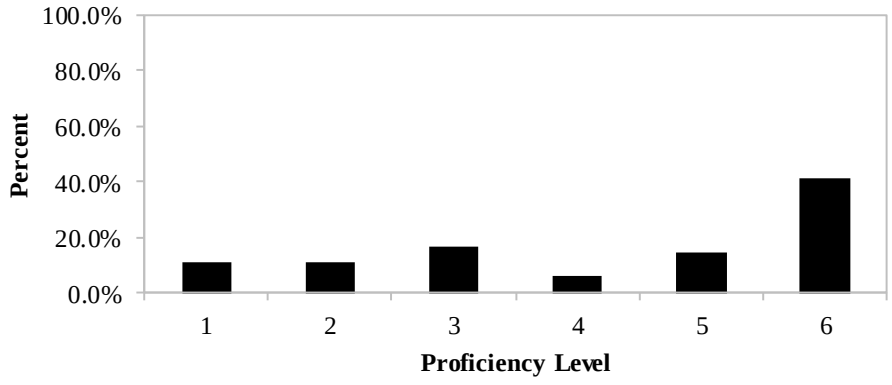
2.5.1.2 *Grades 2–3*

Table 2.5.1.2

Proficiency Level Distribution: List 2-3 S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	23,030	12.52%	17,146	9.38%	40,176	10.96%
2	22,404	12.18%	18,066	9.89%	40,470	11.04%
3	32,675	17.77%	28,177	15.42%	60,852	16.60%
4	10,972	5.97%	10,878	5.95%	21,850	5.96%
5	21,961	11.94%	30,462	16.67%	52,423	14.30%
6	72,847	39.61%	77,985	42.68%	150,832	41.14%
Total	183,889	100.00%	182,714	100.00%	366,603	100.00%

Figure 2.5.1.2
Proficiency Level: List 2-3 S501 Online



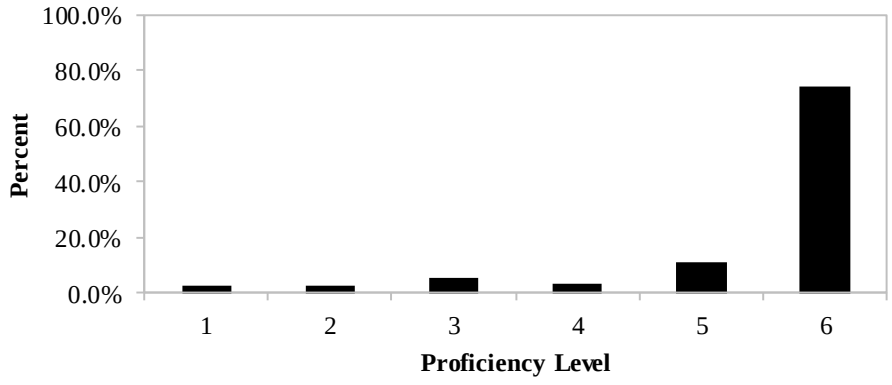
2.5.1.3 *Grades 4–5*

Table 2.5.1.3

Proficiency Level Distribution: List 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	4,043	2.31%	5,174	3.67%	9,217	2.92%
2	3,731	2.14%	4,011	2.84%	7,742	2.45%
3	9,961	5.70%	8,118	5.76%	18,079	5.73%
4	6,607	3.78%	5,001	3.55%	11,608	3.68%
5	19,097	10.93%	16,017	11.36%	35,114	11.12%
6	131,291	75.14%	102,664	72.82%	233,955	74.10%
Total	174,730	100.00%	140,985	100.00%	315,715	100.00%

Figure 2.5.1.3
Proficiency Level: List 4-5 S501 Online

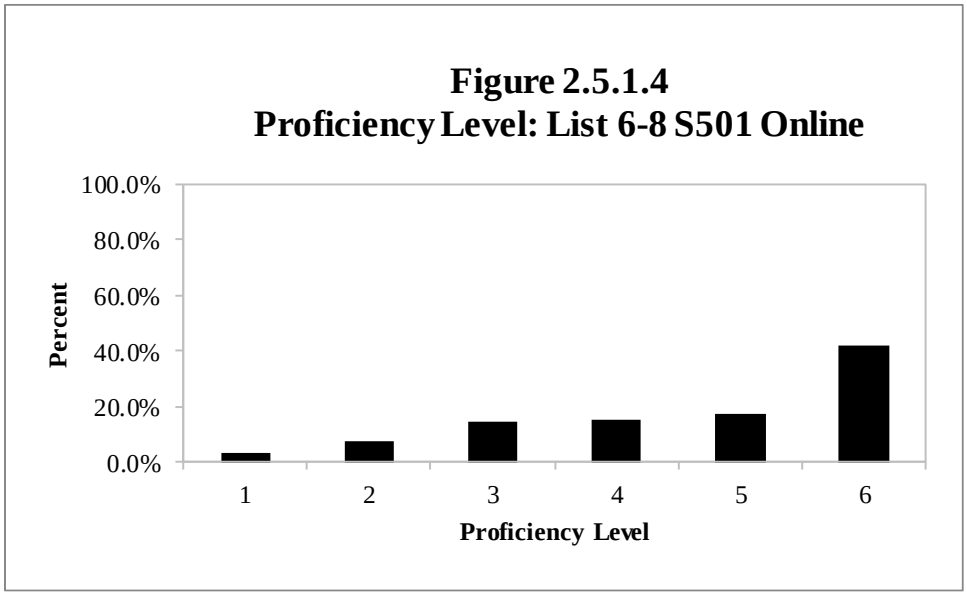


2.5.1.4 *Grades 6–8*

Table 2.5.1.4

Proficiency Level Distribution: List 6-8 S501 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	2,445	2.14%	3,694	3.60%	3,685	4.10%	9,824	3.20%
2	6,158	5.40%	7,468	7.27%	8,760	9.74%	22,386	7.30%
3	15,550	13.64%	15,298	14.90%	13,388	14.89%	44,236	14.43%
4	16,549	14.51%	17,324	16.87%	13,735	15.27%	47,608	15.53%
5	23,157	20.31%	17,672	17.21%	13,376	14.87%	54,205	17.68%
6	50,162	43.99%	41,215	40.14%	36,983	41.13%	128,360	41.86%
Total	114,021	100.00%	102,671	100.00%	89,927	100.00%	306,619	100.00%

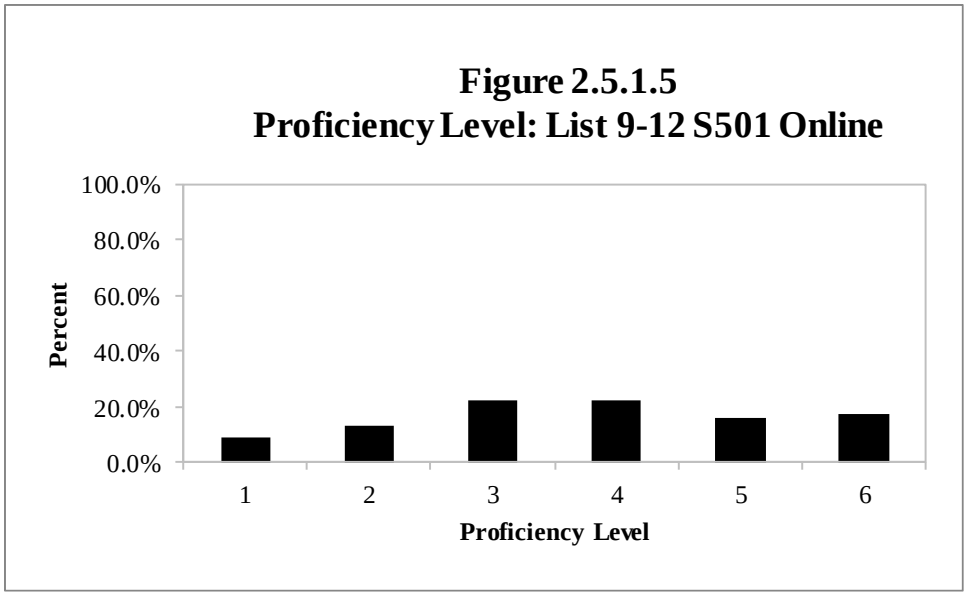


2.5.1.5 *Grades 9–12*

Table 2.5.1.5

Proficiency Level Distribution: List 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	7,709	7.54%	7,105	8.74%	6,059	8.96%	6,760	11.57%	27,633	8.93%
2	15,378	15.04%	10,416	12.81%	8,323	12.31%	5,845	10.01%	39,962	12.91%
3	23,214	22.71%	17,422	21.43%	15,222	22.52%	13,043	22.33%	68,901	22.26%
4	23,049	22.54%	18,027	22.17%	14,502	21.45%	14,472	24.78%	70,050	22.63%
5	15,053	14.72%	14,138	17.39%	11,084	16.40%	9,708	16.62%	49,983	16.15%
6	17,837	17.45%	14,188	17.45%	12,409	18.36%	8,582	14.69%	53,016	17.13%
Total	102,240	100.00%	81,296	100.00%	67,599	100.00%	58,410	100.00%	309,545	100.00%



2.5.2 **Reading**

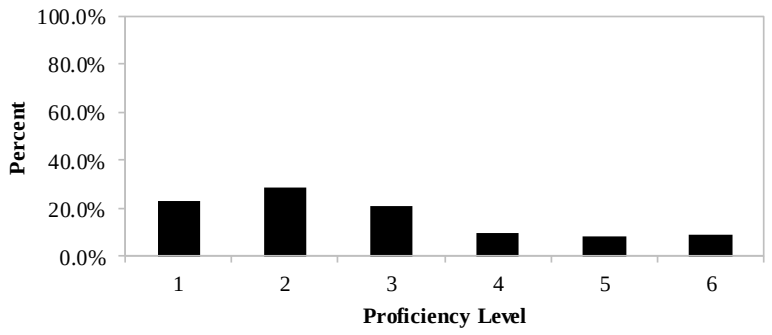
2.5.2.1 *Grade 1*

Table 2.5.2.1

Proficiency Level Distribution: Read 1 S501 Online

	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	40,892	22.75%	40,892	22.75%
2	51,970	28.91%	51,970	28.91%
3	37,854	21.06%	37,854	21.06%
4	17,380	9.67%	17,380	9.67%
5	15,046	8.37%	15,046	8.37%
6	16,597	9.23%	16,597	9.23%
Total	179,739	100.00%	179,739	100.00%

Figure 2.5.2.1
Proficiency Level: Read 1 S501 Online

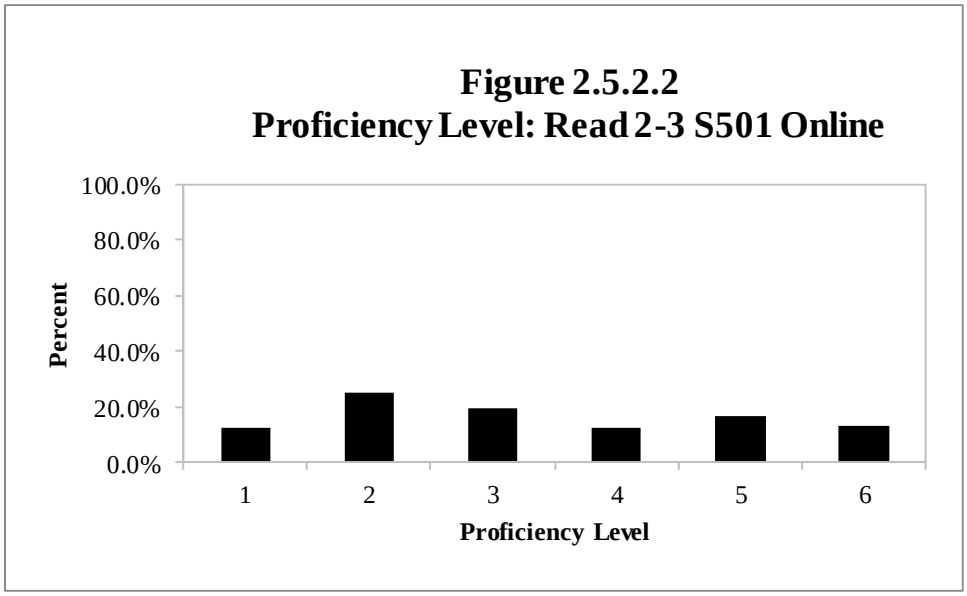


2.5.2.2 *Grades 2–3*

Table 2.5.2.2

Proficiency Level Distribution: Read 2-3 S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	20,317	11.03%	26,344	14.44%	46,661	12.73%
2	45,326	24.61%	46,171	25.30%	91,497	24.96%
3	39,653	21.53%	31,697	17.37%	71,350	19.46%
4	29,979	16.28%	16,997	9.32%	46,976	12.81%
5	28,646	15.56%	32,660	17.90%	61,306	16.72%
6	20,229	10.99%	28,593	15.67%	48,822	13.32%
Total	184,150	100.00%	182,462	100.00%	366,612	100.00%

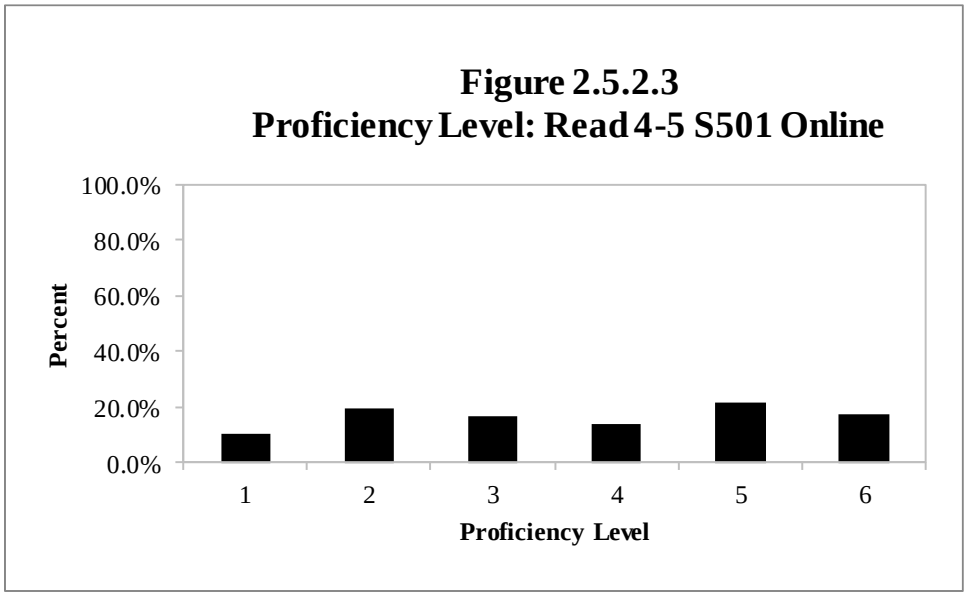


2.5.2.3 *Grades 4–5*

Table 2.5.2.3

Proficiency Level Distribution: Read 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	15,469	9.03%	17,892	12.94%	33,361	10.78%
2	32,720	19.11%	28,008	20.25%	60,728	19.62%
3	24,694	14.42%	26,285	19.00%	50,979	16.47%
4	28,221	16.48%	14,838	10.73%	43,059	13.91%
5	38,073	22.23%	28,596	20.67%	66,669	21.54%
6	32,058	18.72%	22,693	16.41%	54,751	17.69%
Total	171,235	100.00%	138,312	100.00%	309,547	100.00%

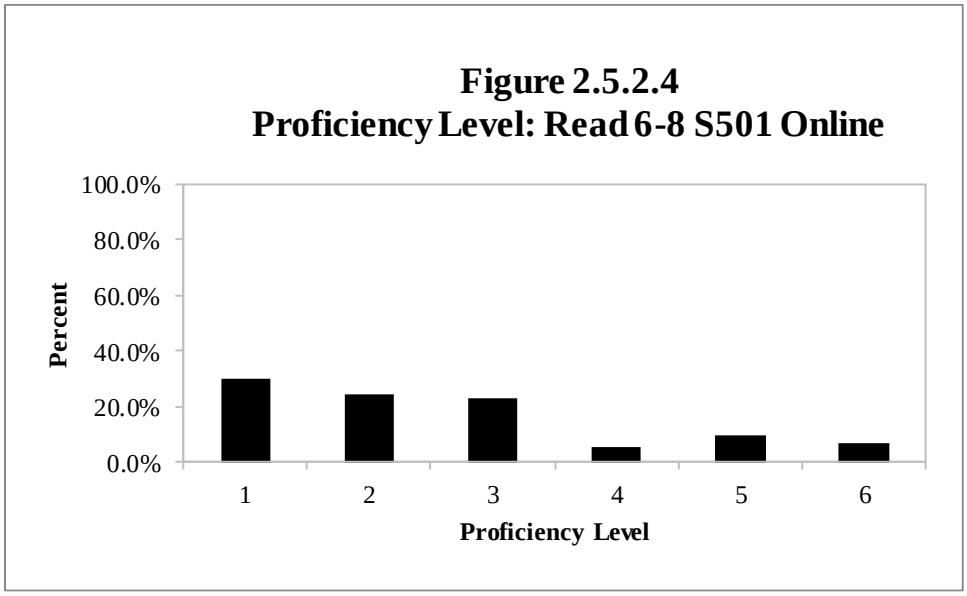


2.5.2.4 *Grades 6–8*

Table 2.5.2.4

Proficiency Level Distribution: Read 6-8 S501 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	33,187	29.29%	30,678	30.10%	28,267	31.81%	92,132	30.30%
2	26,975	23.81%	27,049	26.54%	20,484	23.05%	74,508	24.50%
3	29,500	26.03%	21,495	21.09%	18,497	20.81%	69,492	22.85%
4	7,075	6.24%	4,809	4.72%	4,063	4.57%	15,947	5.24%
5	11,225	9.91%	10,298	10.10%	8,387	9.44%	29,910	9.84%
6	5,348	4.72%	7,585	7.44%	9,169	10.32%	22,102	7.27%
Total	113,310	100.00%	101,914	100.00%	88,867	100.00%	304,091	100.00%

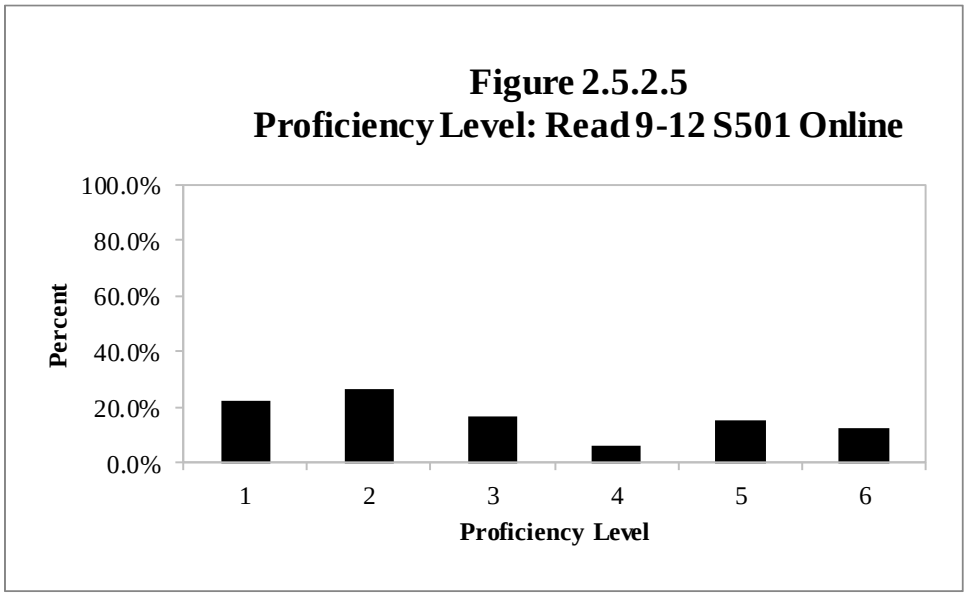


2.5.2.5 *Grades 9–12*

Table 2.5.2.5

Proficiency Level Distribution: Read 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	27,945	27.57%	17,220	21.59%	12,360	18.67%	11,476	19.98%	69,001	22.64%
2	27,383	27.02%	21,149	26.52%	17,140	25.89%	15,745	27.41%	81,417	26.71%
3	16,535	16.31%	13,101	16.43%	11,407	17.23%	9,403	16.37%	50,446	16.55%
4	5,075	5.01%	4,848	6.08%	4,460	6.74%	4,523	7.87%	18,906	6.20%
5	12,904	12.73%	12,236	15.34%	11,340	17.13%	9,666	16.83%	46,146	15.14%
6	11,516	11.36%	11,208	14.05%	9,507	14.36%	6,628	11.54%	38,859	12.75%
Total	101,358	100.00%	79,762	100.00%	66,214	100.00%	57,441	100.00%	304,775	100.00%



2.5.3 **Writing**

2.5.3.1 *Grade 1*

Table 2.5.3.1.1

Proficiency Level Distribution: Writ 1 A S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	41,399	26.13%	41,399	26.13%
2	95,133	60.04%	95,133	60.04%
3	21,823	13.77%	21,823	13.77%
4	104	0.07%	104	0.07%
5	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%
Total	158,459	100.00%	158,459	100.00%

Figure 2.5.3.1.1
Proficiency Level: Writ 1 A S501 Online

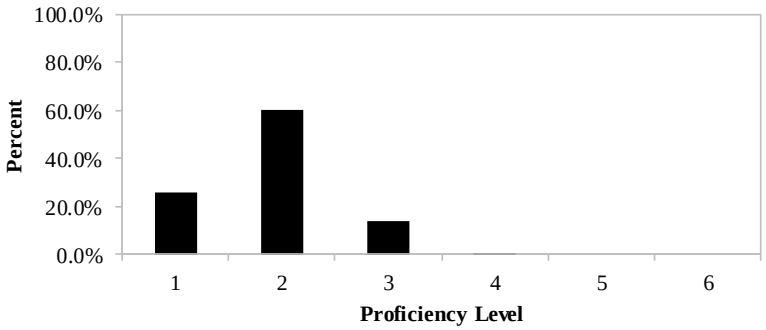


Table 2.5.3.1.2

Proficiency Level Distribution: Writ 1 B/C S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	543	1.91%	543	1.91%
2	6,153	21.67%	6,153	21.67%
3	20,564	72.43%	20,564	72.43%
4	1,099	3.87%	1,099	3.87%
5	32	0.11%	32	0.11%
6	0	0.00%	0	0.00%
Total	28,391	100.00%	28,391	100.00%

Figure 2.5.3.1.2
Proficiency Level: Writ 1 B/C S501 Online

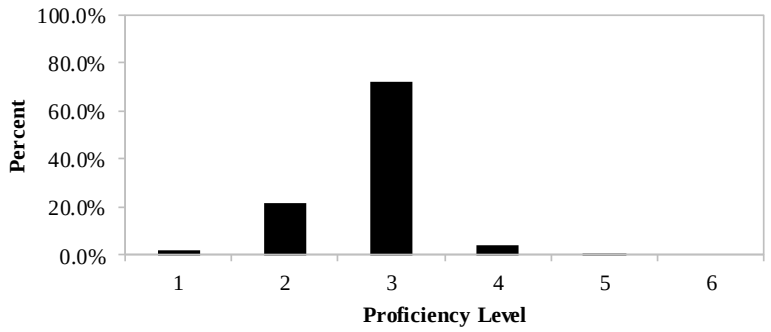
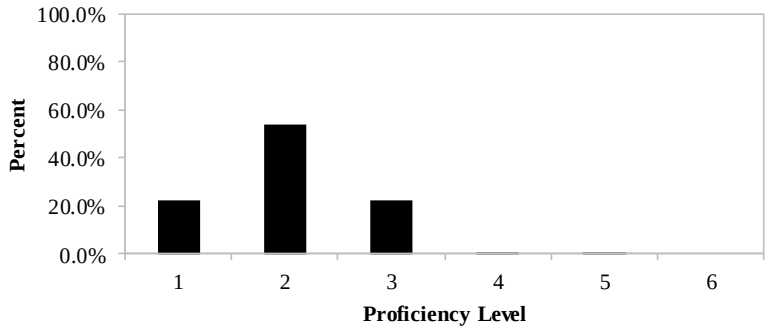


Table 2.5.3.1.3

Proficiency Level Distribution: Writ 1 S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	41,942	22.45%	41,942	22.45%
2	101,286	54.21%	101,286	54.21%
3	42,387	22.69%	42,387	22.69%
4	1,203	0.64%	1,203	0.64%
5	32	0.02%	32	0.02%
6	0	0.00%	0	0.00%
Total	186,850	100.00%	186,850	100.00%

Figure 2.5.3.1.3
Proficiency Level: Writ 1 S501 Online



2.5.3.2 *Grades 2–3*

Table 2.5.3.2.1

Proficiency Level Distribution: Writ 2-3 A S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	16,332	28.66%	11,542	29.85%	27,874	29.14%
2	30,324	53.22%	16,497	42.66%	46,821	48.95%
3	10,159	17.83%	10,575	27.35%	20,734	21.68%
4	161	0.28%	58	0.15%	219	0.23%
5	1	0.00%	0	0.00%	1	0.00%
6	0	0.00%	0	0.00%	0	0.00%
Total	56,977	100.00%	38,672	100.00%	95,649	100.00%

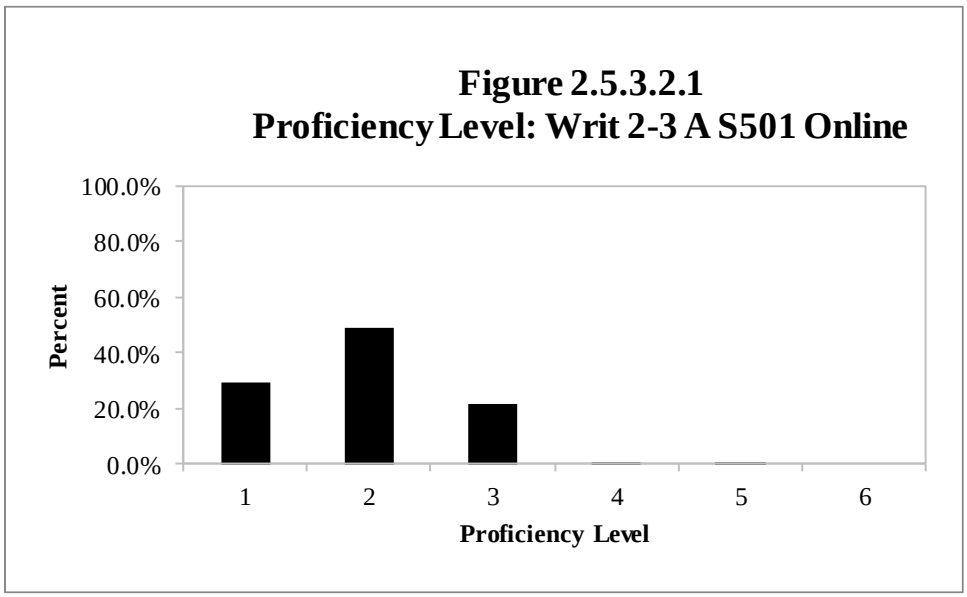


Table 2.5.3.2.2

Proficiency Level Distribution: Writ 2-3 B/C S501 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	2,414	1.76%	656	0.43%	3,070	1.06%
2	14,591	10.64%	5,304	3.46%	19,895	6.85%
3	110,112	80.29%	114,693	74.80%	224,805	77.39%
4	9,998	7.29%	32,414	21.14%	42,412	14.60%
5	34	0.02%	263	0.17%	297	0.10%
6	2	0.00%	7	0.00%	9	0.00%
Total	137,151	100.00%	153,337	100.00%	290,488	100.00%

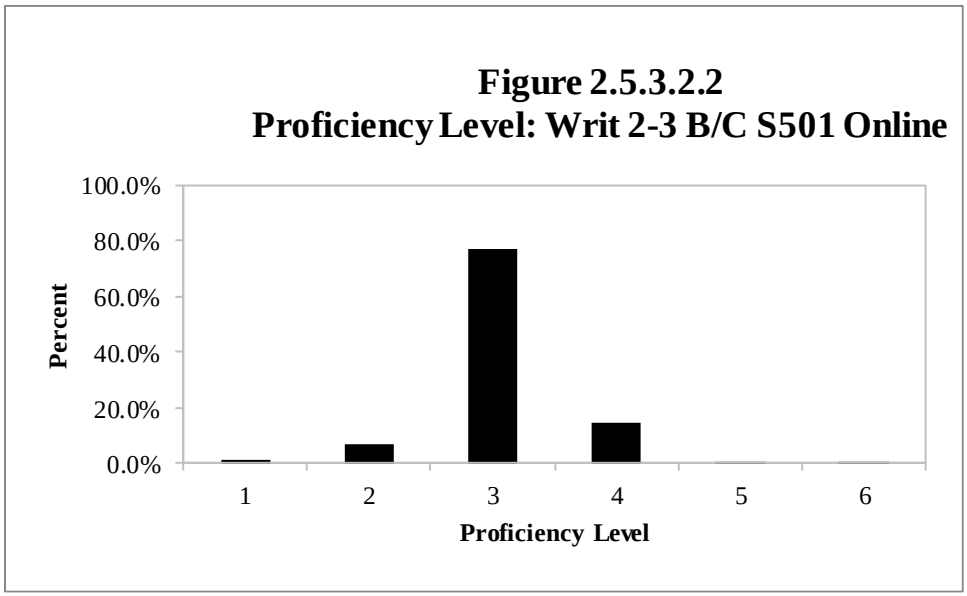
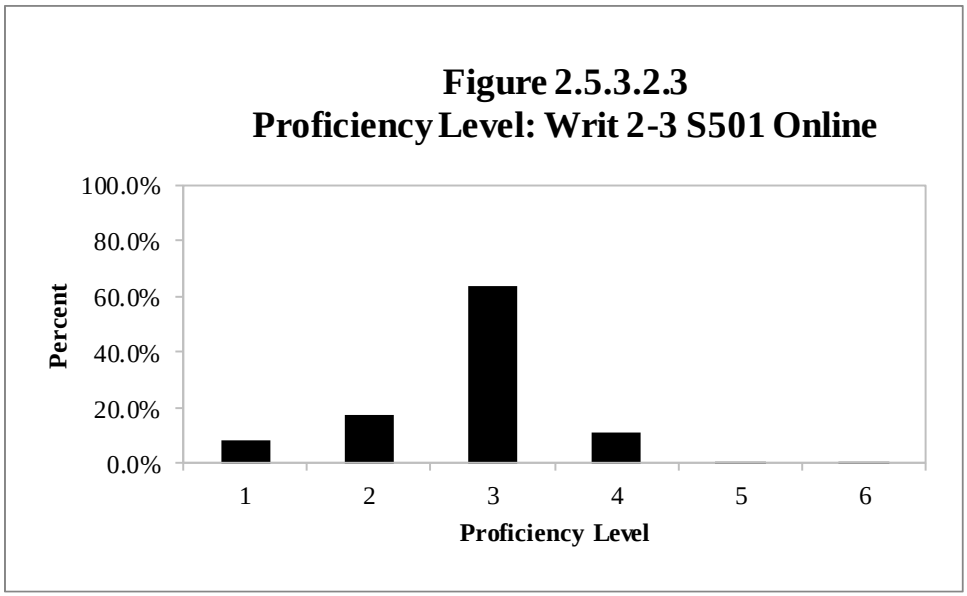


Table 2.5.3.2.3

Proficiency Level Distribution: Writ 2-3 S501 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	18,746	9.66%	12,198	6.35%	30,944	8.01%
2	44,915	23.14%	21,801	11.35%	66,716	17.28%
3	120,271	61.95%	125,268	65.24%	245,539	63.59%
4	10,159	5.23%	32,472	16.91%	42,631	11.04%
5	35	0.02%	263	0.14%	298	0.08%
6	2	0.00%	7	0.00%	9	0.00%
Total	194,128	100.00%	192,009	100.00%	386,137	100.00%



2.5.3.3 *Grades 4–5*

Table 2.5.3.3.1

Proficiency Level Distribution: Writ 4-5 A S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	8,317	33.29%	6,384	25.61%	14,701	29.45%
2	7,676	30.72%	7,137	28.63%	14,813	29.68%
3	8,826	35.33%	11,025	44.23%	19,851	39.77%
4	164	0.66%	381	1.53%	545	1.09%
5	1	0.00%	1	0.00%	2	0.00%
6	0	0.00%	0	0.00%	0	0.00%
Total	24,984	100.00%	24,928	100.00%	49,912	100.00%

Figure 2.5.3.3.1
Proficiency Level: Writ 4-5 A S501 Online

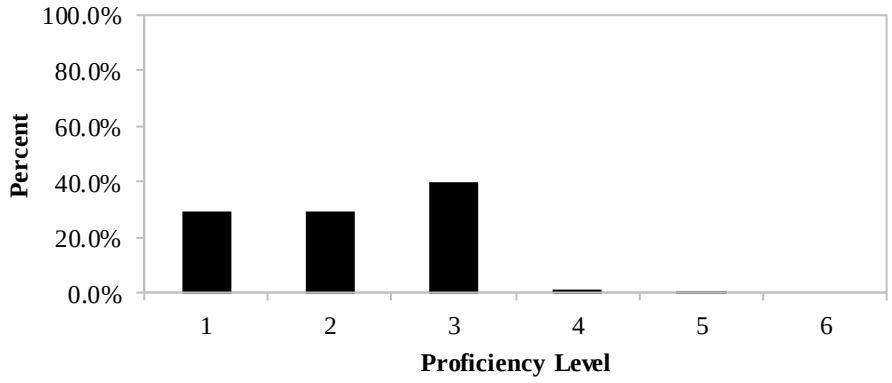


Table 2.5.3.3.2

Proficiency Level Distribution: Writ 4-5 B/C S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	954	0.63%	376	0.32%	1,330	0.50%
2	1,929	1.28%	609	0.52%	2,538	0.95%
3	84,680	56.13%	49,272	41.92%	133,952	49.91%
4	59,475	39.42%	61,261	52.12%	120,736	44.98%
5	2,940	1.95%	5,608	4.77%	8,548	3.18%
6	895	0.59%	414	0.35%	1,309	0.49%
Total	150,873	100.00%	117,540	100.00%	268,413	100.00%

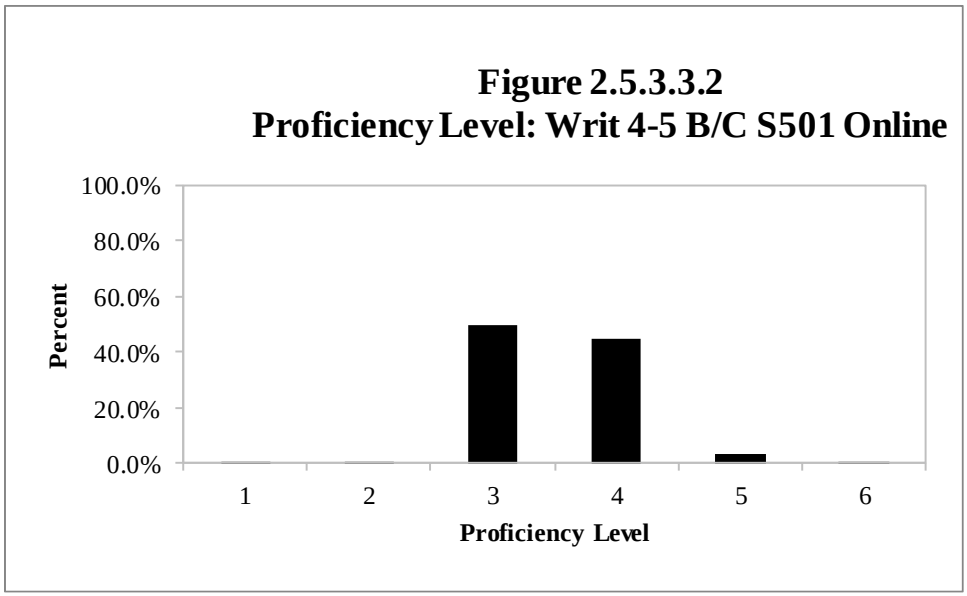
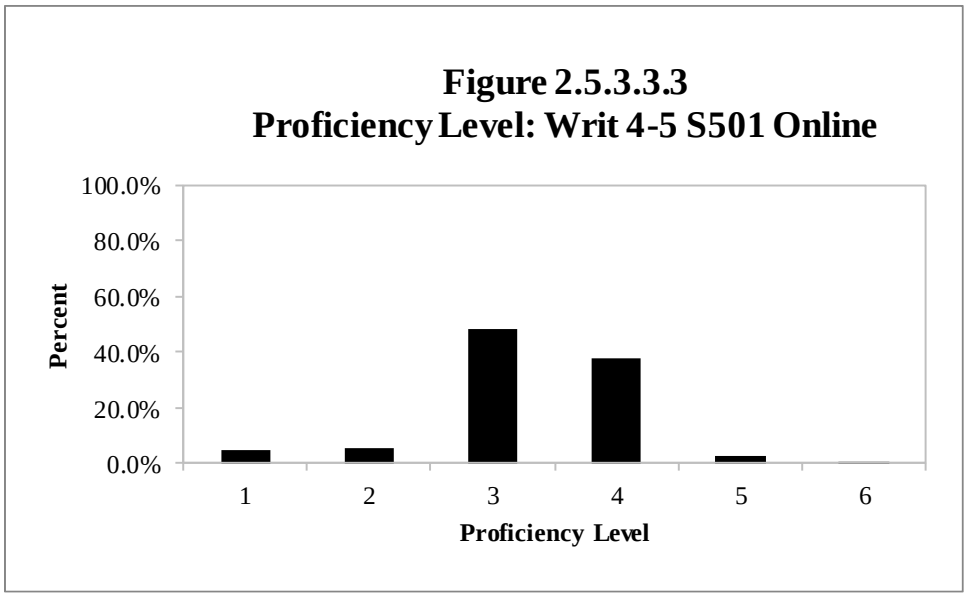


Table 2.5.3.3.3

Proficiency Level Distribution: Writ 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	9,271	5.27%	6,760	4.74%	16,031	5.04%
2	9,605	5.46%	7,746	5.44%	17,351	5.45%
3	93,506	53.17%	60,297	42.32%	153,803	48.32%
4	59,639	33.91%	61,642	43.27%	121,281	38.10%
5	2,941	1.67%	5,609	3.94%	8,550	2.69%
6	895	0.51%	414	0.29%	1,309	0.41%
Total	175,857	100.00%	142,468	100.00%	318,325	100.00%



2.5.3.4 *Grades 6–8*

Table 2.5.3.4.1

Proficiency Level Distribution: Writ 6-8 A S501 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	6,766	20.12%	8,764	22.35%	10,328	27.72%	25,858	23.48%
2	10,789	32.08%	16,944	43.21%	12,440	33.38%	40,173	36.48%
3	15,865	47.17%	13,374	34.11%	14,315	38.42%	43,554	39.55%
4	214	0.64%	132	0.34%	180	0.48%	526	0.48%
5	0	0.00%	0	0.00%	0	0.00%	0	0.00%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	33,634	100.00%	39,214	100.00%	37,263	100.00%	110,111	100.00%

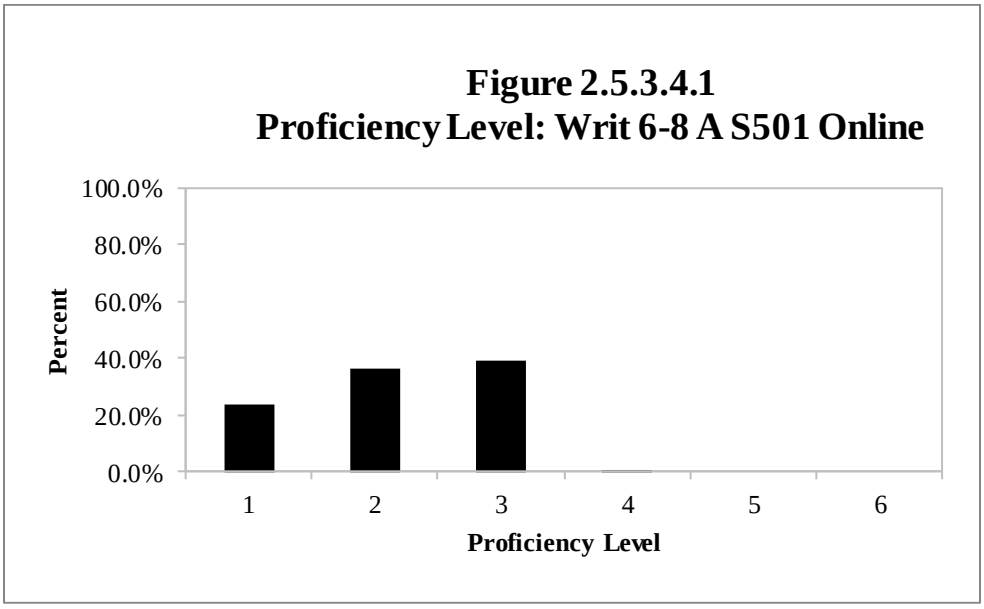


Table 2.5.3.4.2

Proficiency Level Distribution: Writ 6-8 B/C S501 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	610	0.74%	360	0.55%	409	0.76%	1,379	0.68%
2	7,123	8.60%	7,155	10.93%	3,925	7.31%	18,203	9.01%
3	60,255	72.75%	38,779	59.25%	39,893	74.29%	138,927	68.78%
4	14,712	17.76%	19,107	29.19%	9,316	17.35%	43,135	21.36%
5	119	0.14%	48	0.07%	145	0.27%	312	0.15%
6	5	0.01%	2	0.00%	10	0.02%	17	0.01%
Total	82,824	100.00%	65,451	100.00%	53,698	100.00%	201,973	100.00%

Figure 2.5.3.4.2
Proficiency Level: Writ 6-8 B/C S501 Online

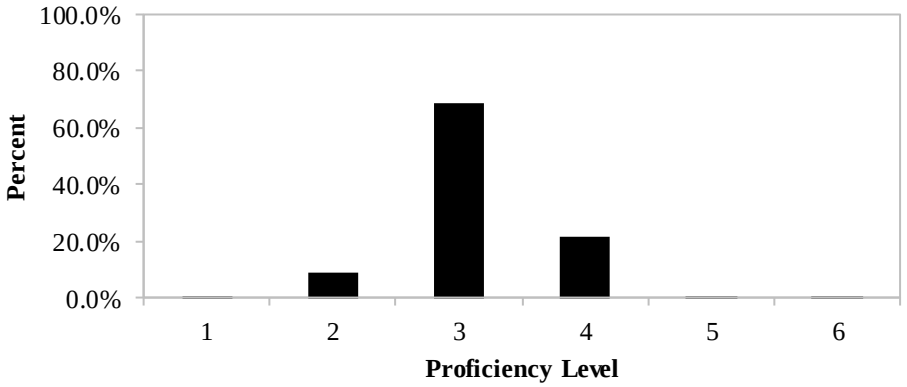
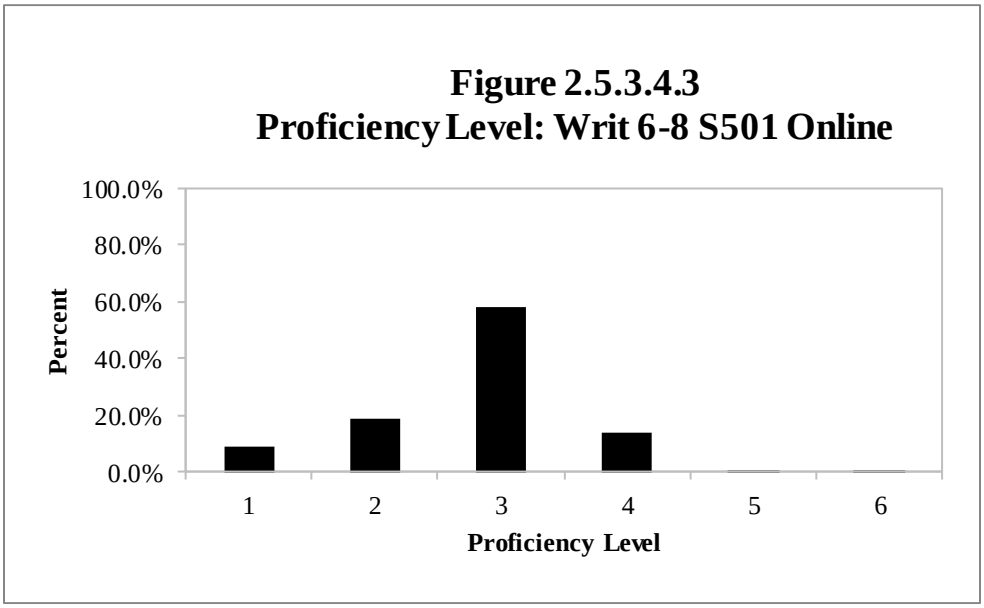


Table 2.5.3.4.3

Proficiency Level Distribution: Writ 6-8 S501 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	7,376	6.33%	9,124	8.72%	10,737	11.80%	27,237	8.73%
2	17,912	15.38%	24,099	23.02%	16,365	17.99%	58,376	18.71%
3	76,120	65.36%	52,153	49.83%	54,208	59.59%	182,481	58.47%
4	14,926	12.82%	19,239	18.38%	9,496	10.44%	43,661	13.99%
5	119	0.10%	48	0.05%	145	0.16%	312	0.10%
6	5	0.00%	2	0.00%	10	0.01%	17	0.01%
Total	116,458	100.00%	104,665	100.00%	90,961	100.00%	312,084	100.00%



2.5.3.5 *Grades 9–12*

Table 2.5.3.5.1
Proficiency Level Distribution: Writ 9-12 A S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	13,313	28.09%	6,826	23.01%	5,518	26.15%	5,598	34.99%	31,255	27.38%
2	14,577	30.75%	7,167	24.16%	5,827	27.61%	2,499	15.62%	30,070	26.34%
3	15,050	31.75%	14,266	48.09%	8,461	40.09%	7,537	47.11%	45,314	39.69%
4	4,436	9.36%	1,382	4.66%	1,294	6.13%	363	2.27%	7,475	6.55%
5	23	0.05%	23	0.08%	5	0.02%	3	0.02%	54	0.05%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	47,399	100.00%	29,664	100.00%	21,105	100.00%	16,000	100.00%	114,168	100.00%

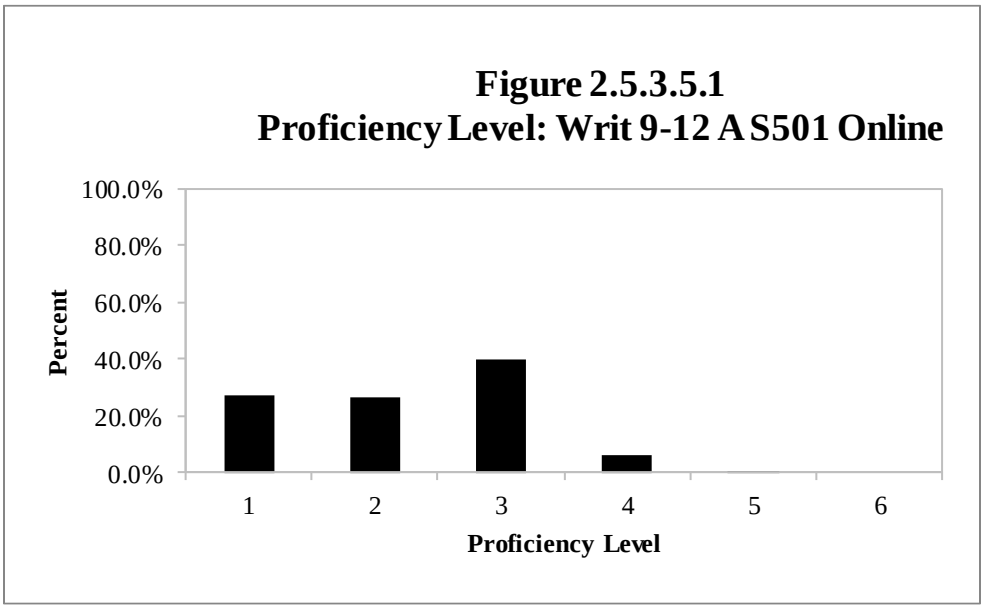


Table 2.5.3.5.2

Proficiency Level Distribution: Writ 9-12 B/C S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	532	0.92%	786	1.46%	1,045	2.17%	1,770	4.05%	4,133	2.03%
2	2,474	4.26%	4,516	8.40%	8,293	17.21%	6,722	15.38%	22,005	10.80%
3	35,775	61.55%	38,675	71.92%	28,530	59.22%	25,853	59.16%	128,833	63.22%
4	19,030	32.74%	9,435	17.54%	9,900	20.55%	9,264	21.20%	47,629	23.37%
5	299	0.51%	365	0.68%	406	0.84%	90	0.21%	1,160	0.57%
6	11	0.02%	1	0.00%	1	0.00%	0	0.00%	13	0.01%
Total	58,121	100.00%	53,778	100.00%	48,175	100.00%	43,699	100.00%	203,773	100.00%

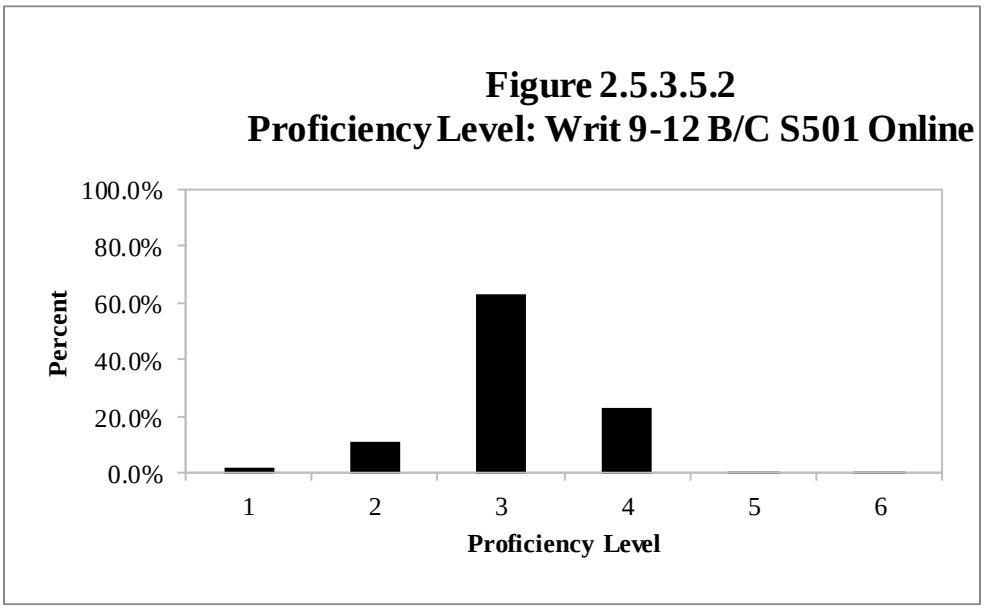
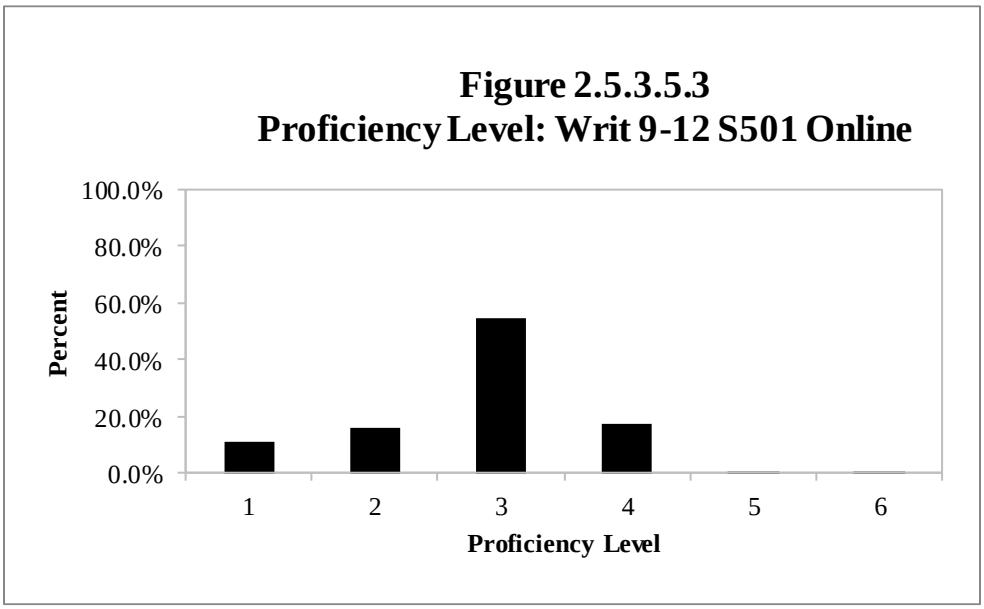


Table 2.5.3.5.3

Proficiency Level Distribution: Writ 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	13,845	13.12%	7,612	9.12%	6,563	9.47%	7,368	12.34%	35,388	11.13%
2	17,051	16.16%	11,683	14.00%	14,120	20.38%	9,221	15.45%	52,075	16.38%
3	50,825	48.17%	52,941	63.45%	36,991	53.39%	33,390	55.93%	174,147	54.77%
4	23,466	22.24%	10,817	12.96%	11,194	16.16%	9,627	16.13%	55,104	17.33%
5	322	0.31%	388	0.46%	411	0.59%	93	0.16%	1,214	0.38%
6	11	0.01%	1	0.00%	1	0.00%	0	0.00%	13	0.00%
Total	105,520	100.00%	83,442	100.00%	69,280	100.00%	59,699	100.00%	317,941	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 S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	7,109	100.00%	7,109	100.00%
Total	7,109	100.00%	7,109	100.00%

Figure 2.5.4.1.1
Proficiency Level: Spek 1 Pre-A S501 Online

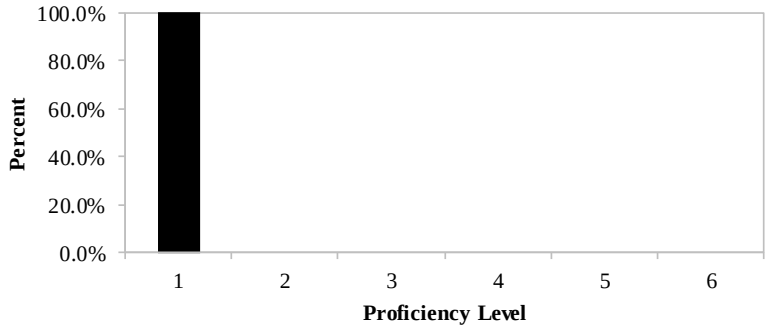


Table 2.5.4.1.2

Proficiency Level Distribution: Spek 1 A S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	16,856	24.84%	16,856	24.84%
2	32,270	47.55%	32,270	47.55%
3	15,221	22.43%	15,221	22.43%
4	3,234	4.77%	3,234	4.77%
5	283	0.42%	283	0.42%
6	0	0.00%	0	0.00%
Total	67,864	100.00%	67,864	100.00%

Figure 2.5.4.1.2
Proficiency Level: Spek 1 A S501 Online

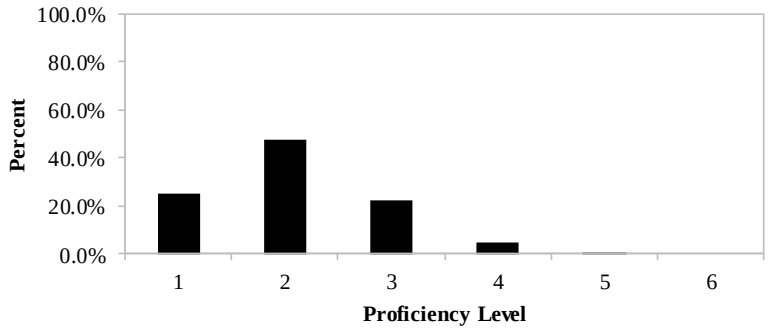


Table 2.5.4.1.3

Proficiency Level Distribution: Spek 1 B/C S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	4,915	4.92%	4,915	4.92%
2	27,694	27.72%	27,694	27.72%
3	43,839	43.88%	43,839	43.88%
4	21,814	21.83%	21,814	21.83%
5	1,560	1.56%	1,560	1.56%
6	88	0.09%	88	0.09%
Total	99,910	100.00%	99,910	100.00%

Figure 2.5.4.1.3
Proficiency Level: Spek 1 B/C S501 Online

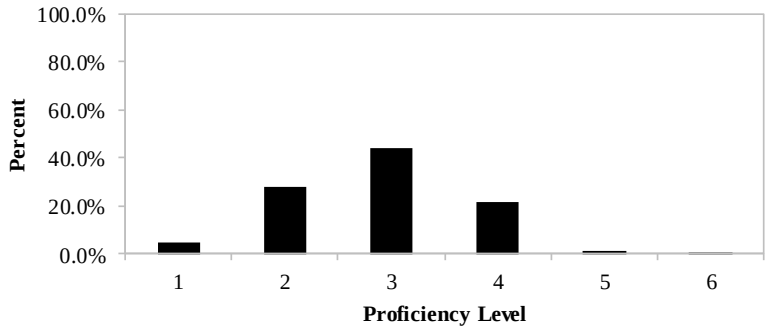
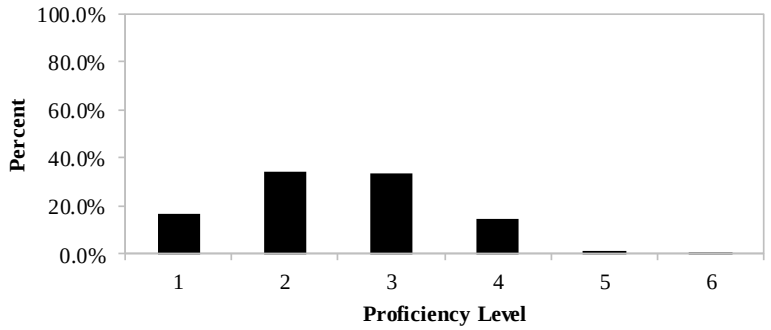


Table 2.5.4.1.4

Proficiency Level Distribution: Spek 1 S501 Online

Level	Grade 1		Total	
	Count	Percent	Count	Percent
1	28,880	16.51%	28,880	16.51%
2	59,964	34.29%	59,964	34.29%
3	59,060	33.77%	59,060	33.77%
4	25,048	14.32%	25,048	14.32%
5	1,843	1.05%	1,843	1.05%
6	88	0.05%	88	0.05%
Total	174,883	100.00%	174,883	100.00%

Figure 2.5.4.1.4
Proficiency Level: Spek 1 S501 Online



2.5.4.2 *Grades 2–3*

Table 2.5.4.2.1

Proficiency Level Distribution: Spek 2-3 Pre-A S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	7,246	100.00%	9,858	100.00%	17,104	100.00%
Total	7,246	100.00%	9,858	100.00%	17,104	100.00%

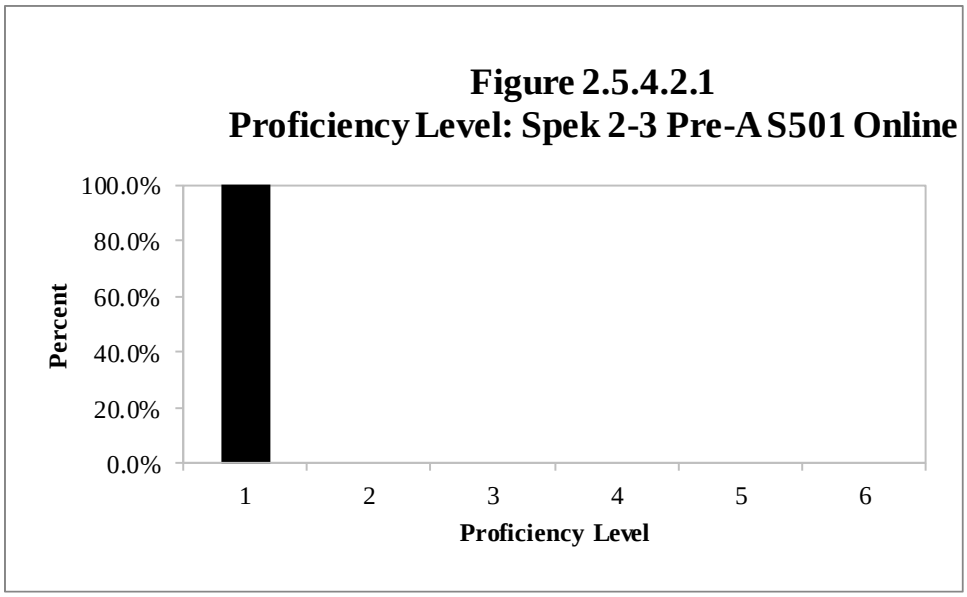


Table 2.5.4.2.2

Proficiency Level Distribution: Spek 2-3 A S501 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	12,725	29.03%	11,071	28.89%	23,796	28.96%
2	19,392	44.23%	19,366	50.54%	38,758	47.18%
3	9,887	22.55%	6,697	17.48%	16,584	20.19%
4	1,807	4.12%	1,122	2.93%	2,929	3.57%
5	28	0.06%	62	0.16%	90	0.11%
6	0	0.00%	0	0.00%	0	0.00%
Total	43,839	100.00%	38,318	100.00%	82,157	100.00%

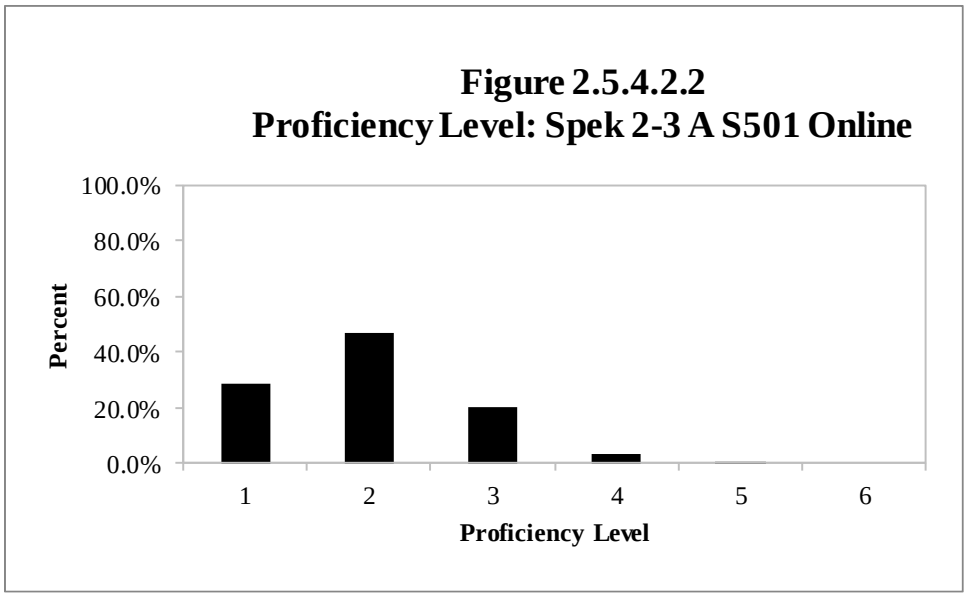


Table 2.5.4.2.3

Proficiency Level Distribution: Spek 2-3 B/C S501 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	5,967	4.54%	3,815	2.86%	9,782	3.69%
2	48,746	37.12%	27,789	20.82%	76,535	28.90%
3	61,241	46.64%	83,565	62.59%	144,806	54.68%
4	14,357	10.93%	17,322	12.97%	31,679	11.96%
5	911	0.69%	734	0.55%	1,645	0.62%
6	97	0.07%	279	0.21%	376	0.14%
Total	131,319	100.00%	133,504	100.00%	264,823	100.00%

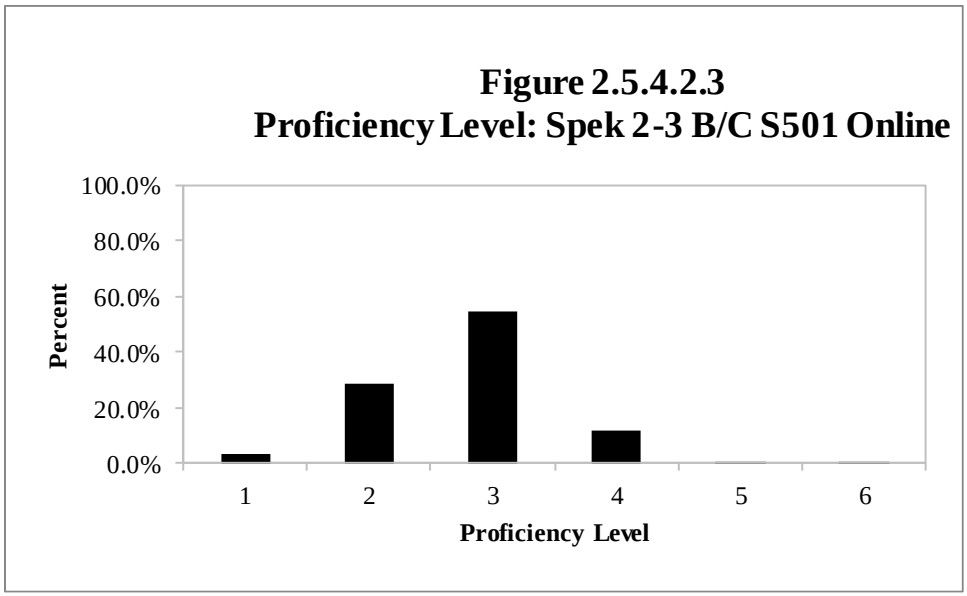
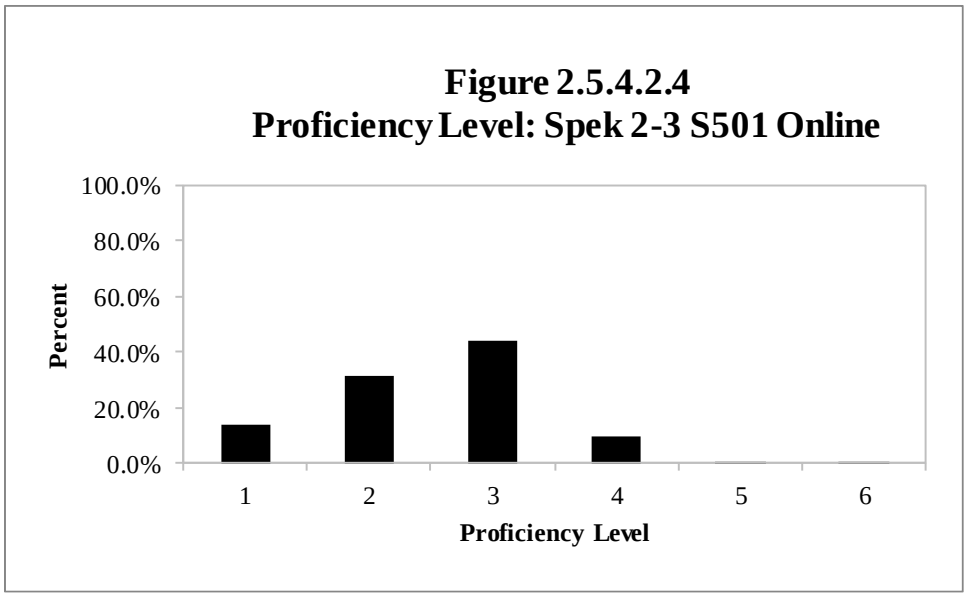


Table 2.5.4.2.4

Proficiency Level Distribution: Spek 2-3 S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	25,938	14.22%	24,744	13.62%	50,682	13.92%
2	68,138	37.36%	47,155	25.95%	115,293	31.67%
3	71,128	38.99%	90,262	49.68%	161,390	44.33%
4	16,164	8.86%	18,444	10.15%	34,608	9.51%
5	939	0.51%	796	0.44%	1,735	0.48%
6	97	0.05%	279	0.15%	376	0.10%
Total	182,404	100.00%	181,680	100.00%	364,084	100.00%



2.5.4.3 *Grades 4–5*

Table 2.5.4.3.1

Proficiency Level Distribution: Spek 4-5 Pre-A S501 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	2,303	100.00%	4,067	100.00%	6,370	100.00%
Total	2,303	100.00%	4,067	100.00%	6,370	100.00%

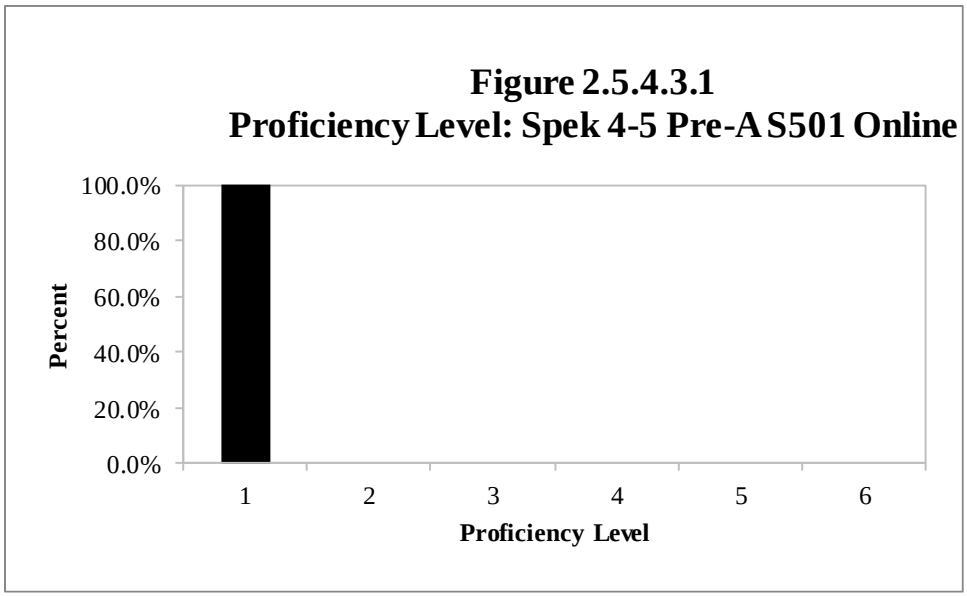


Table 2.5.4.3.2

Proficiency Level Distribution: Spek 4-5 A S501 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	6,810	38.29%	6,291	45.31%	13,101	41.37%
2	7,752	43.59%	4,564	32.87%	12,316	38.89%
3	2,949	16.58%	2,765	19.92%	5,714	18.04%
4	263	1.48%	262	1.89%	525	1.66%
5	11	0.06%	2	0.01%	13	0.04%
6	0	0.00%	0	0.00%	0	0.00%
Total	17,785	100.00%	13,884	100.00%	31,669	100.00%

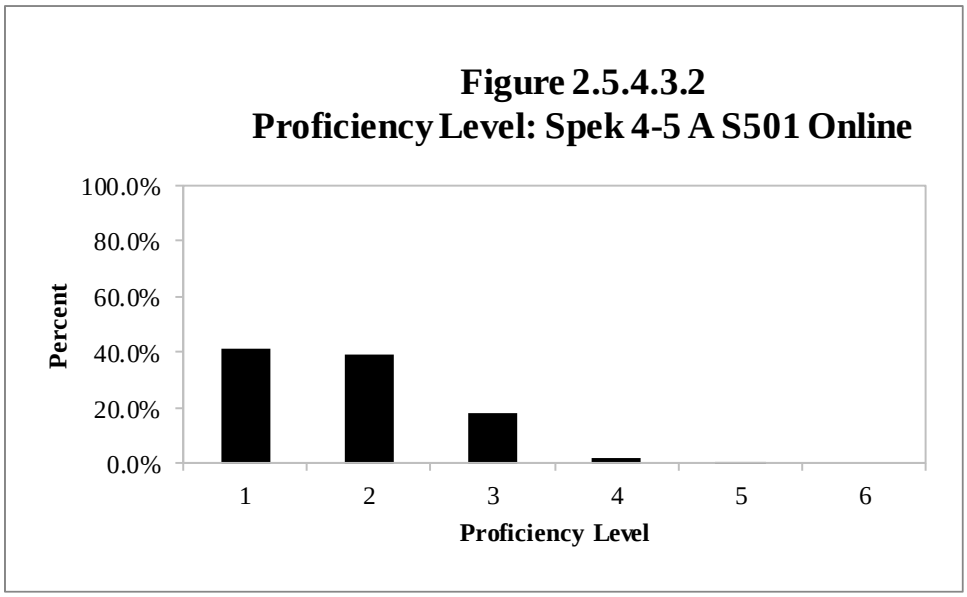


Table 2.5.4.3.3

Proficiency Level Distribution: Spek 4-5 B/C S501 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	4,196	2.77%	4,954	4.09%	9,150	3.35%
2	34,503	22.75%	23,713	19.58%	58,216	21.34%
3	71,889	47.40%	56,535	46.69%	128,424	47.08%
4	38,868	25.63%	33,630	27.77%	72,498	26.58%
5	1,933	1.27%	2,174	1.80%	4,107	1.51%
6	274	0.18%	83	0.07%	357	0.13%
Total	151,663	100.00%	121,089	100.00%	272,752	100.00%

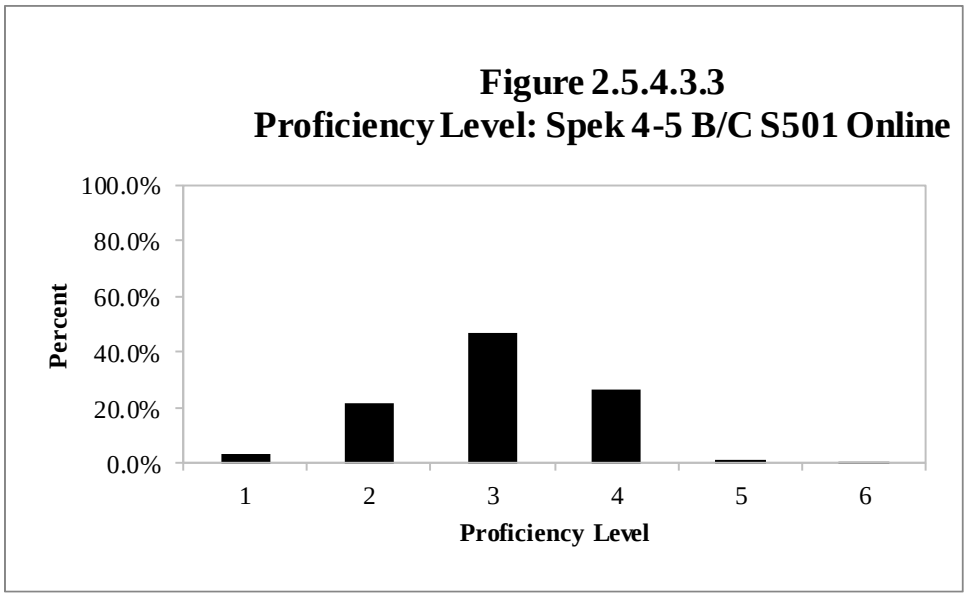
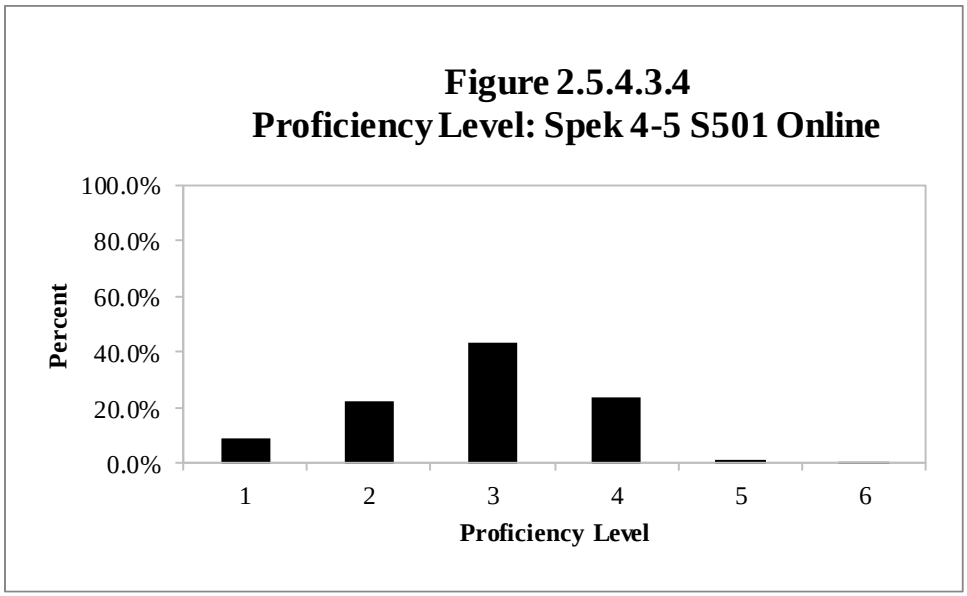


Table 2.5.4.3.4

Proficiency Level Distribution: Spek 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	13,309	7.75%	15,312	11.01%	28,621	9.21%
2	42,255	24.60%	28,277	20.34%	70,532	22.69%
3	74,838	43.57%	59,300	42.65%	134,138	43.16%
4	39,131	22.78%	33,892	24.38%	73,023	23.50%
5	1,944	1.13%	2,176	1.57%	4,120	1.33%
6	274	0.16%	83	0.06%	357	0.11%
Total	171,751	100.00%	139,040	100.00%	310,791	100.00%



2.5.4.4 *Grades 6–8*

Table 2.5.4.4.1

Proficiency Level Distribution: Spek 6-8 Pre-A S501 Online

Level	Grade 6		Grade 7		Grade 8		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	2,259	100.00%	3,570	100.00%	3,704	100.00%	9,533	100.00%
Total	2,259	100.00%	3,570	100.00%	3,704	100.00%	9,533	100.00%

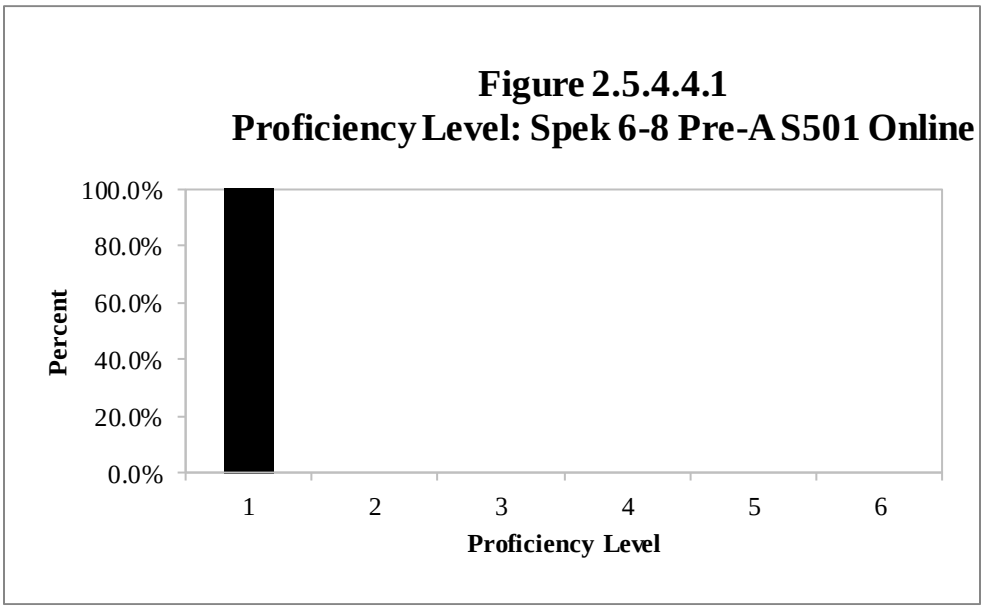


Table 2.5.4.4.2

Proficiency Level Distribution: Spek 6-8 A S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	8,292	43.54%	7,431	46.18%	13,401	49.47%	29,124	46.80%
2	8,071	42.38%	6,621	41.14%	7,548	27.87%	22,240	35.74%
3	2,481	13.03%	1,901	11.81%	5,941	21.93%	10,323	16.59%
4	202	1.06%	139	0.86%	191	0.71%	532	0.85%
5	0	0.00%	0	0.00%	6	0.02%	6	0.01%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	19,046	100.00%	16,092	100.00%	27,087	100.00%	62,225	100.00%

Figure 2.5.4.4.2
Proficiency Level: Spek 6-8 A S501 Online

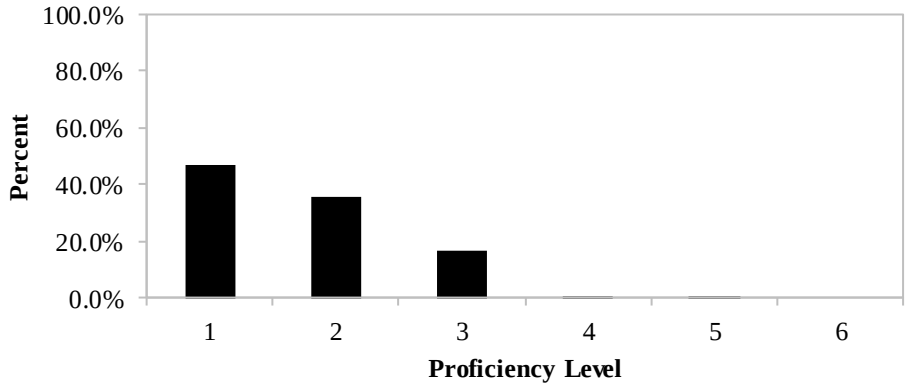


Table 2.5.4.4.3

Proficiency Level Distribution: Spek 6-8 B/C S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	5,280	5.68%	7,511	9.02%	5,330	8.95%	18,121	7.68%
2	24,951	26.83%	17,973	21.58%	13,993	23.51%	56,917	24.14%
3	47,636	51.22%	47,556	57.09%	28,671	48.17%	123,863	52.52%
4	14,686	15.79%	10,012	12.02%	11,088	18.63%	35,786	15.18%
5	436	0.47%	233	0.28%	411	0.69%	1,080	0.46%
6	11	0.01%	16	0.02%	27	0.05%	54	0.02%
Total	93,000	100.00%	83,301	100.00%	59,520	100.00%	235,821	100.00%

Figure 2.5.4.4.3
Proficiency Level: Spek 6-8 B/C S501 Online

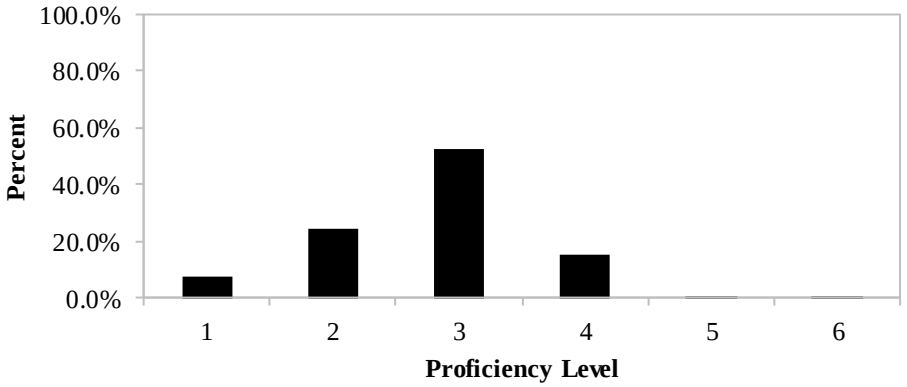
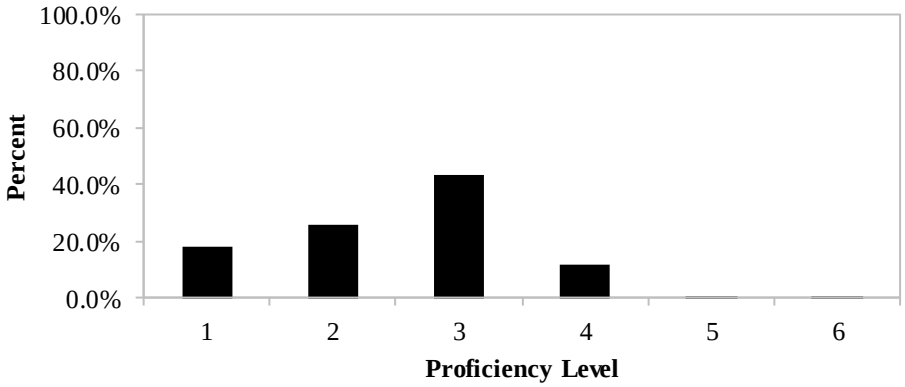


Table 2.5.4.4.4

Proficiency Level Distribution: Spek 6-8 S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	15,831	13.85%	18,512	17.98%	22,435	24.84%	56,778	18.46%
2	33,022	28.89%	24,594	23.89%	21,541	23.85%	79,157	25.74%
3	50,117	43.84%	49,457	48.03%	34,612	38.33%	134,186	43.63%
4	14,888	13.02%	10,151	9.86%	11,279	12.49%	36,318	11.81%
5	436	0.38%	233	0.23%	417	0.46%	1,086	0.35%
6	11	0.01%	16	0.02%	27	0.03%	54	0.02%
Total	114,305	100.00%	102,963	100.00%	90,311	100.00%	307,579	100.00%

Figure 2.5.4.4.4
Proficiency Level: Spek 6-8 S501 Online



2.5.4.5 *Grades 9–12*

Table 2.5.4.5.1

Proficiency Level Distribution: Spek 9-12 Pre-A S501 Online

Level	Grade 9		Grade 10		Grade 11		Grade 12		Total	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent
1	6,238	100.00%	5,280	100.00%	4,105	100.00%	4,266	100.00%	19,889	100.00%
Total	6,238	100.00%	5,280	100.00%	4,105	100.00%	4,266	100.00%	19,889	100.00%

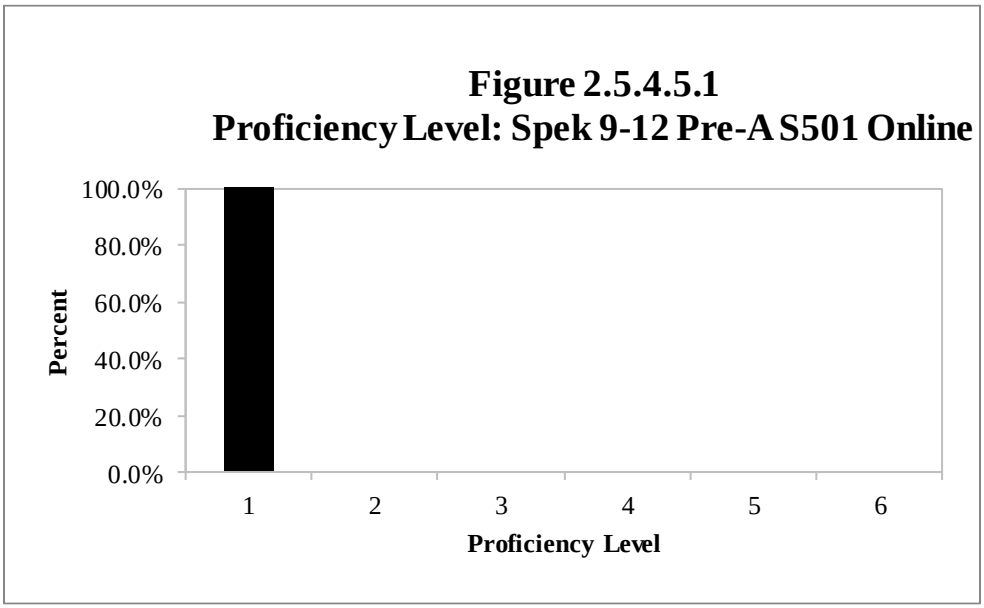


Table 2.5.4.5.2

Proficiency Level Distribution: Spek 9-12 A S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	31,060	53.65%	15,733	48.45%	7,322	54.22%	8,222	32.79%	62,337	48.34%
2	12,993	22.44%	8,145	25.08%	5,088	37.68%	11,857	47.28%	38,083	29.53%
3	13,388	23.13%	8,311	25.59%	1,023	7.58%	4,874	19.44%	27,596	21.40%
4	439	0.76%	286	0.88%	71	0.53%	123	0.49%	919	0.71%
5	11	0.02%	0	0.00%	0	0.00%	0	0.00%	11	0.01%
6	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Total	57,891	100.00%	32,475	100.00%	13,504	100.00%	25,076	100.00%	128,946	100.00%

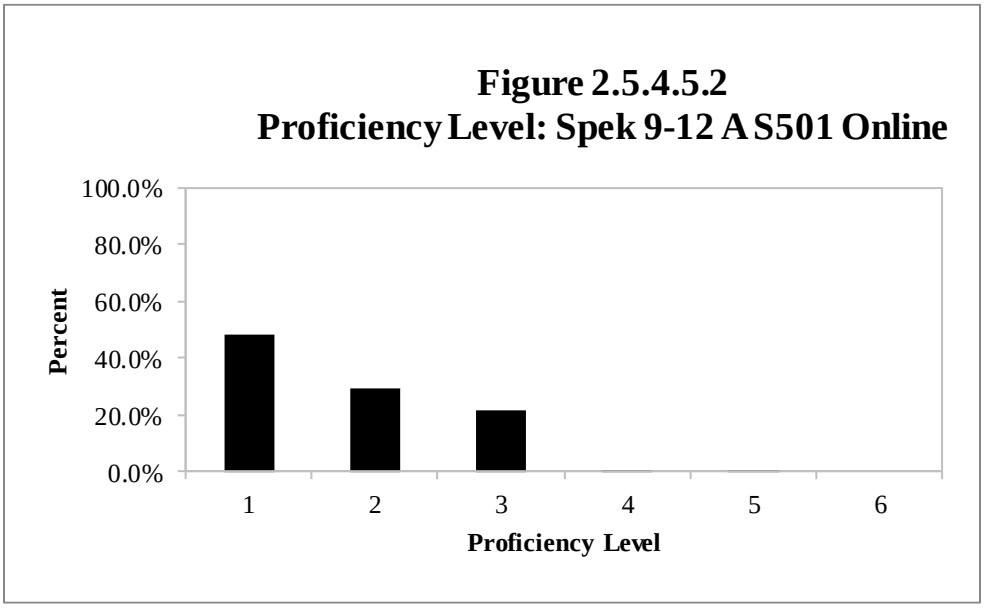


Table 2.5.4.5.3

Proficiency Level Distribution: Spek 9-12 B/C S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	3,603	9.27%	4,284	9.88%	8,276	16.70%	3,565	12.33%	19,728	12.27%
2	8,808	22.66%	9,581	22.09%	15,713	31.71%	7,746	26.79%	41,848	26.04%
3	22,890	58.90%	27,177	62.65%	22,864	46.14%	16,478	56.98%	89,409	55.63%
4	3,508	9.03%	2,256	5.20%	2,615	5.28%	1,044	3.61%	9,423	5.86%
5	36	0.09%	54	0.12%	56	0.11%	63	0.22%	209	0.13%
6	19	0.05%	28	0.06%	31	0.06%	22	0.08%	100	0.06%
Total	38,864	100.00%	43,380	100.00%	49,555	100.00%	28,918	100.00%	160,717	100.00%

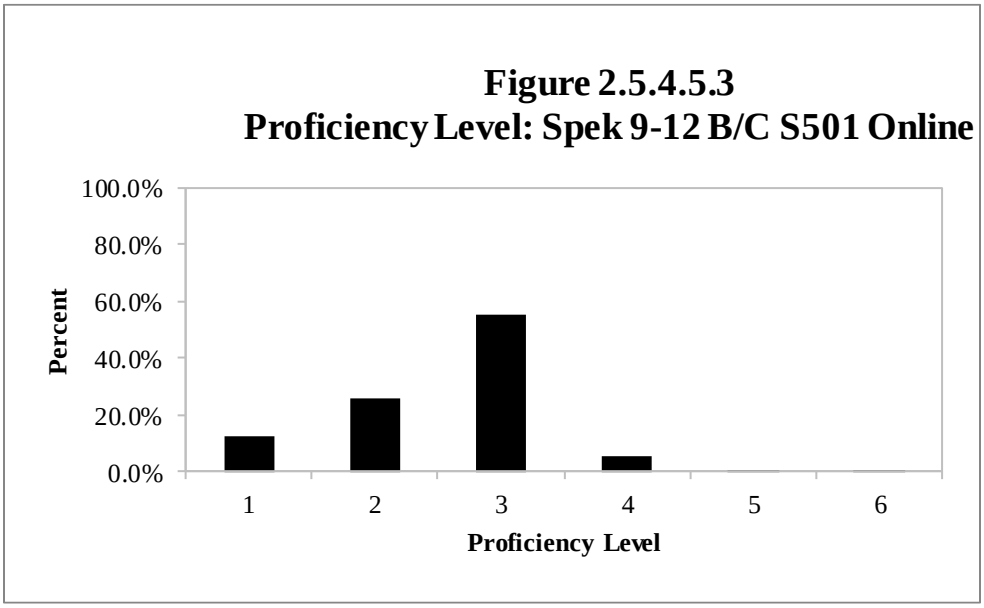
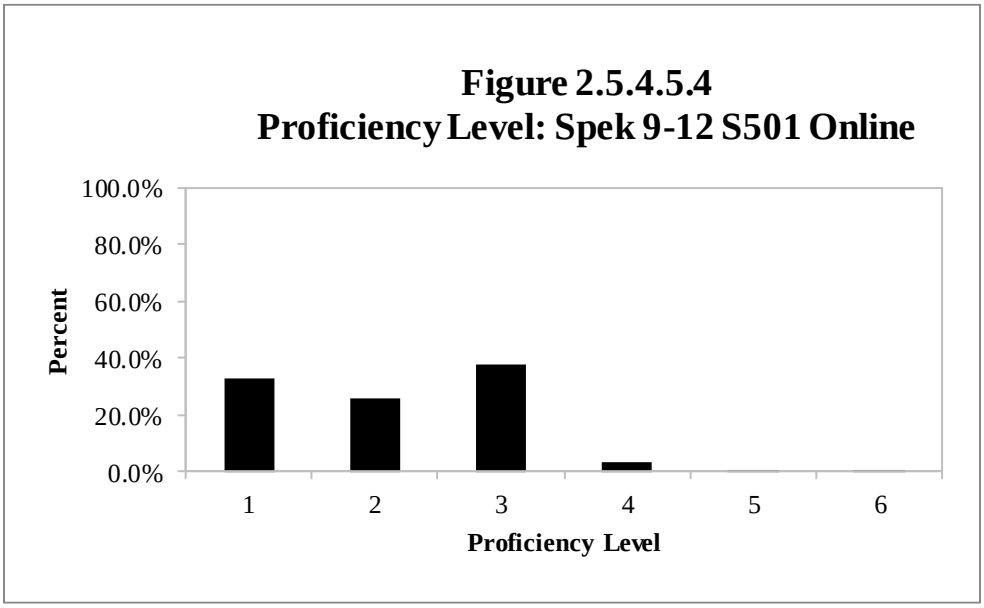


Table 2.5.4.5.4

Proficiency Level Distribution: Spek 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	40,901	39.71%	25,297	31.18%	19,703	29.34%	16,053	27.55%	101,954	32.94%
2	21,801	21.17%	17,726	21.85%	20,801	30.97%	19,603	33.65%	79,931	25.82%
3	36,278	35.22%	35,488	43.74%	23,887	35.57%	21,352	36.65%	117,005	37.80%
4	3,947	3.83%	2,542	3.13%	2,686	4.00%	1,167	2.00%	10,342	3.34%
5	47	0.05%	54	0.07%	56	0.08%	63	0.11%	220	0.07%
6	19	0.02%	28	0.03%	31	0.05%	22	0.04%	100	0.03%
Total	102,993	100.00%	81,135	100.00%	67,164	100.00%	58,260	100.00%	309,552	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 of 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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	111	256	1.0
1	200	45	1.7
2	214	32	1.8
3	223	28	1.8
4	231	27	1.9
5	239	28	2.0
6	247	31	2.2
7	258	35	2.5
8	271	39	2.8
9	287	41	3.1
10	305	42	3.4
11	323	42	3.7
12	340	40	4.0
13	355	38	4.4
14	368	36	4.6
15	381	36	4.9
16	395	40	5.5
17	415	52	6.0
18	447	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	111	256	1.0
1	214	45	1.8
2	228	32	1.9
3	237	28	1.9
4	244	27	2.1
5	252	28	2.3
6	261	31	2.6
7	271	35	2.8
8	285	39	3.1
9	301	41	3.4
10	318	42	3.6
11	336	42	3.9
12	353	40	4.3
13	368	38	4.6
14	382	36	5.0
15	395	36	5.5
16	409	40	6.0
17	429	52	6.0
18	460	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	133	237	1.0	1.0
1	203	45	1.6	1.6
2	217	33	1.7	1.7
3	226	29	1.8	1.8
4	234	28	1.9	1.8
5	242	28	2.0	1.9
6	251	31	2.2	2.1
7	261	34	2.5	2.3
8	275	39	2.8	2.7
9	291	41	3.1	3.1
10	308	42	3.4	3.3
11	326	42	3.7	3.6
12	343	40	4.0	3.9
13	358	38	4.3	4.2
14	372	36	4.6	4.5
15	385	36	4.9	4.8
16	399	40	5.4	5.2
17	419	52	6.0	6.0
18	450	94	6.0	6.0

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

Table 2.6.3.2.2

Raw Score to Scale Score to Proficiency Level Conversion: Writ 2-3 B/C S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	133	256	1.0	1.0
1	211	45	1.7	1.6
2	226	35	1.8	1.8
3	237	32	1.9	1.9
4	247	31	2.1	2.0
5	257	30	2.4	2.2
6	266	31	2.6	2.5
7	276	34	2.9	2.8
8	289	38	3.1	3.0
9	304	41	3.4	3.3
10	321	42	3.6	3.6
11	339	41	3.9	3.8
12	355	40	4.2	4.1
13	370	38	4.6	4.5
14	384	37	4.9	4.7
15	398	38	5.4	5.1
16	414	42	6.0	5.8
17	435	53	6.0	6.0
18	467	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	155	256	1.0	1.0
1	237	45	1.7	1.7
2	251	33	1.8	1.8
3	260	29	1.9	1.9
4	268	28	2.0	2.0
5	276	28	2.4	2.3
6	284	31	2.8	2.6
7	295	35	3.1	3.0
8	308	39	3.3	3.2
9	324	41	3.5	3.4
10	342	42	3.8	3.7
11	359	42	4.1	4.0
12	376	40	4.5	4.3
13	391	38	4.8	4.6
14	405	36	5.1	4.9
15	418	36	5.7	5.4
16	432	40	6.0	5.9
17	452	52	6.0	6.0
18	484	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	155	256	1.0	1.0
1	272	45	2.2	2.1
2	286	32	2.9	2.7
3	295	28	3.1	3.0
4	302	27	3.2	3.1
5	310	28	3.3	3.2
6	319	31	3.4	3.4
7	329	35	3.6	3.5
8	343	39	3.8	3.7
9	359	41	4.1	4.0
10	376	42	4.5	4.3
11	394	42	4.8	4.7
12	411	40	5.4	5.1
13	426	38	6.0	5.7
14	440	36	6.0	6.0
15	453	36	6.0	6.0
16	467	40	6.0	6.0
17	487	52	6.0	6.0
18	518	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	188	158	1.2	1.1	1.0
1	238	45	1.7	1.6	1.5
2	253	33	1.8	1.7	1.6
3	262	29	1.9	1.8	1.7
4	271	28	2.1	1.9	1.8
5	279	29	2.3	2.1	1.9
6	288	31	2.6	2.4	2.2
7	298	34	3.0	2.7	2.5
8	312	38	3.2	3.1	3.0
9	327	41	3.4	3.3	3.2
10	345	42	3.7	3.6	3.5
11	362	42	4.0	3.9	3.8
12	379	40	4.3	4.2	4.1
13	394	38	4.6	4.5	4.4
14	408	37	4.9	4.7	4.6
15	421	37	5.2	5.0	4.9
16	436	40	5.8	5.5	5.3
17	456	52	6.0	6.0	5.9
18	488	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	188	184	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	276	27	2.2	2.0	1.9
5	284	28	2.5	2.3	2.1
6	292	31	2.8	2.5	2.3
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	460	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	232	86	1.3	1.2	1.1	1.0
1	256	45	1.6	1.5	1.4	1.2
2	270	32	1.7	1.6	1.5	1.4
3	280	28	1.9	1.7	1.6	1.5
4	287	27	1.9	1.8	1.7	1.6
5	295	28	2.2	1.9	1.8	1.7
6	304	31	2.5	2.2	1.9	1.8
7	314	35	2.8	2.5	2.2	1.9
8	328	39	3.1	3.0	2.7	2.3
9	344	41	3.4	3.3	3.1	3.0
10	361	42	3.7	3.5	3.4	3.3
11	379	42	4.0	3.8	3.7	3.6
12	396	40	4.3	4.2	4.1	3.9
13	411	38	4.6	4.5	4.4	4.2
14	425	36	4.9	4.7	4.6	4.5
15	438	36	5.2	5.0	4.9	4.8
16	452	40	5.5	5.3	5.2	5.0
17	472	52	6.0	5.8	5.6	5.4
18	503	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	232	93	1.3	1.2	1.1	1.0
1	260	45	1.6	1.5	1.4	1.3
2	274	33	1.8	1.7	1.6	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	308	31	2.6	2.3	2.0	1.8
7	319	34	3.0	2.7	2.4	2.0
8	332	38	3.2	3.1	2.8	2.5
9	348	41	3.4	3.3	3.2	3.0
10	365	42	3.7	3.6	3.5	3.3
11	383	42	4.0	3.9	3.8	3.7
12	400	40	4.4	4.2	4.1	4.0
13	415	38	4.7	4.5	4.4	4.3
14	429	36	4.9	4.8	4.7	4.6
15	442	37	5.3	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	508	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	106	48	1.0
1	106	48	1.0
2	119	40	1.1
3	133	37	1.2
4	146	40	1.4
5	159	48	1.5
6	172	61	1.6

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

Table 2.6.4.1.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 1 A S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
0	106	44	1.0
1	106	44	1.0
2	115	39	1.0
3	127	35	1.2
4	138	33	1.3
5	148	34	1.4
6	159	36	1.5
7	172	38	1.6
8	185	39	1.7
9	199	40	1.9
10	214	43	2.1
11	233	49	2.5
12	258	55	2.9
13	284	52	3.4
14	307	48	3.9
15	327	47	4.3
16	347	50	4.7
17	367	58	5.1
18	387	73	5.6

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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G1
6	106	42	1.0
7	160	31	1.5
8	168	31	1.6
9	177	30	1.7
10	185	30	1.7
11	193	29	1.8
12	201	30	1.9
13	210	30	2.0
14	218	31	2.2
15	227	33	2.3
16	238	35	2.5
17	249	37	2.7
18	262	38	3.0
19	276	38	3.3
20	289	37	3.5
21	300	35	3.7
22	311	34	4.0
23	321	33	4.2
24	331	33	4.4
25	341	34	4.6
26	352	36	4.8
27	365	39	5.0
28	378	44	5.4
29	391	51	5.7
30	404	60	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	118	38	1.0	1.0
1	118	38	1.0	1.0
2	118	38	1.0	1.0
3	127	37	1.1	1.0
4	140	40	1.2	1.1
5	153	48	1.3	1.3
6	166	61	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
0	118	36	1.0	1.0
1	118	36	1.0	1.0
2	118	36	1.0	1.0
3	122	35	1.0	1.0
4	132	34	1.1	1.1
5	143	35	1.2	1.2
6	155	37	1.3	1.3
7	168	39	1.5	1.4
8	182	39	1.6	1.5
9	196	40	1.7	1.6
10	212	43	1.9	1.8
11	230	49	2.1	1.9
12	255	55	2.6	2.4
13	282	52	3.1	2.9
14	304	47	3.6	3.4
15	324	46	4.0	3.8
16	345	50	4.4	4.2
17	366	59	4.8	4.6
18	387	76	5.3	5.0

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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G2	PL for G3
6	118	33	1.0	1.0
7	154	31	1.3	1.3
8	163	32	1.4	1.3
9	172	31	1.5	1.4
10	181	31	1.6	1.5
11	190	31	1.7	1.6
12	198	31	1.7	1.6
13	207	31	1.8	1.7
14	216	32	1.9	1.8
15	225	33	2.0	1.9
16	236	34	2.3	2.0
17	247	36	2.5	2.2
18	259	37	2.7	2.5
19	272	37	2.9	2.7
20	285	36	3.2	3.0
21	296	35	3.4	3.2
22	308	34	3.7	3.5
23	318	34	3.9	3.7
24	329	34	4.1	3.9
25	339	35	4.3	4.1
26	351	36	4.5	4.3
27	364	39	4.8	4.5
28	377	44	5.0	4.8
29	390	51	5.3	5.1
30	425	81	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	130	42	1.0	1.0
1	130	42	1.0	1.0
2	133	41	1.0	1.0
3	147	39	1.1	1.1
4	161	41	1.3	1.2
5	175	49	1.4	1.3
6	189	62	1.5	1.4

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

Table 2.6.4.3.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 4-5 A S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
0	130	40	1.0	1.0
1	130	40	1.0	1.0
2	130	40	1.0	1.0
3	143	37	1.1	1.1
4	155	36	1.2	1.1
5	168	38	1.3	1.2
6	182	41	1.4	1.4
7	198	42	1.6	1.5
8	214	42	1.7	1.6
9	231	43	1.8	1.7
10	248	45	2.0	1.9
11	268	49	2.4	2.2
12	292	53	2.9	2.7
13	317	51	3.4	3.3
14	339	48	3.9	3.7
15	360	48	4.3	4.1
16	383	52	4.7	4.5
17	406	61	5.2	4.9
18	429	79	5.8	5.6

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

Table 2.6.4.3.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 4-5 B/C S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G4	PL for G5
6	130	40	1.0	1.0
7	190	37	1.5	1.4
8	202	36	1.6	1.5
9	214	34	1.7	1.6
10	224	33	1.8	1.7
11	234	32	1.9	1.8
12	243	31	1.9	1.8
13	252	31	2.1	1.9
14	261	32	2.3	2.0
15	270	33	2.5	2.2
16	281	34	2.7	2.5
17	292	36	2.9	2.7
18	304	37	3.2	3.0
19	317	37	3.4	3.3
20	330	37	3.7	3.5
21	341	35	3.9	3.8
22	353	35	4.2	4.0
23	363	34	4.3	4.2
24	374	34	4.5	4.4
25	385	35	4.7	4.6
26	396	36	4.9	4.8
27	409	39	5.3	5.0
28	422	44	5.6	5.4
29	435	51	6.0	5.7
30	448	60	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	148	44	1.0	1.0	1.0
1	148	44	1.0	1.0	1.0
2	155	40	1.1	1.0	1.0
3	168	37	1.2	1.1	1.1
4	182	40	1.3	1.2	1.2
5	196	49	1.4	1.4	1.3
6	210	64	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
0	148	41	1.0	1.0	1.0
1	148	41	1.0	1.0	1.0
2	152	39	1.1	1.0	1.0
3	165	35	1.2	1.1	1.1
4	176	35	1.3	1.2	1.2
5	187	37	1.3	1.3	1.2
6	201	40	1.4	1.4	1.3
7	216	42	1.6	1.5	1.5
8	232	41	1.7	1.6	1.6
9	247	41	1.8	1.7	1.7
10	263	43	1.9	1.8	1.8
11	282	49	2.3	2.1	1.9
12	307	55	2.9	2.7	2.5
13	334	52	3.4	3.3	3.2
14	356	48	3.9	3.7	3.6
15	376	46	4.2	4.1	3.9
16	396	50	4.6	4.4	4.3
17	416	58	4.9	4.8	4.6
18	436	73	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 S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G6	PL for G7	PL for G8
6	148	41	1.0	1.0	1.0
7	207	36	1.5	1.4	1.4
8	219	35	1.6	1.5	1.5
9	229	33	1.7	1.6	1.5
10	239	31	1.7	1.7	1.6
11	248	31	1.8	1.7	1.7
12	256	30	1.9	1.8	1.7
13	265	30	1.9	1.9	1.8
14	273	31	2.1	1.9	1.9
15	283	33	2.3	2.1	1.9
16	293	35	2.5	2.4	2.2
17	305	37	2.8	2.7	2.5
18	318	38	3.1	3.0	2.8
19	331	38	3.4	3.2	3.1
20	344	37	3.6	3.5	3.3
21	356	35	3.9	3.7	3.6
22	367	34	4.1	3.9	3.8
23	377	33	4.2	4.1	4.0
24	387	33	4.4	4.3	4.1
25	397	34	4.6	4.5	4.3
26	408	36	4.8	4.6	4.5
27	421	39	5.1	4.9	4.7
28	434	44	5.5	5.2	5.0
29	447	51	5.8	5.6	5.4
30	463	63	6.0	6.0	6.0

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

2.6.4.5 *Grades 9–12*

Table 2.6.4.5.1

Raw Score to Scale Score to Proficiency Level Conversion: Spek 9-12 Pre-A S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	172	38	1.1	1.0	1.0	1.0
1	172	38	1.1	1.0	1.0	1.0
2	172	38	1.1	1.0	1.0	1.0
3	180	37	1.1	1.1	1.1	1.0
4	194	40	1.2	1.2	1.2	1.1
5	208	49	1.3	1.3	1.3	1.2
6	222	64	1.5	1.4	1.4	1.3

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

Table 2.6.4.5.2

Raw Score to Scale Score to Proficiency Level Conversion: Spek 9-12 A S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
0	172	36	1.1	1.0	1.0	1.0
1	172	36	1.1	1.0	1.0	1.0
2	172	36	1.1	1.0	1.0	1.0
3	175	35	1.1	1.1	1.0	1.0
4	186	34	1.2	1.1	1.1	1.1
5	197	35	1.3	1.2	1.2	1.1
6	209	37	1.4	1.3	1.3	1.2
7	222	39	1.5	1.4	1.4	1.3
8	236	40	1.6	1.5	1.5	1.4
9	251	40	1.7	1.6	1.6	1.6
10	266	43	1.8	1.7	1.7	1.7
11	285	49	1.9	1.9	1.8	1.8
12	310	55	2.5	2.3	2.2	2.2
13	336	52	3.1	3.0	2.9	2.8
14	359	48	3.5	3.4	3.3	3.2
15	379	47	3.8	3.7	3.6	3.5
16	400	50	4.2	4.1	4.0	3.9
17	421	59	4.6	4.5	4.4	4.3
18	442	76	5.0	4.9	4.8	4.7

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

Table 2.6.4.5.3

Raw Score to Scale Score to Proficiency Level Conversion: Spek 9-12 B/C S501 Online

Raw Score	Scale Score	CSEM x 1.96	PL for G9	PL for G10	PL for G11	PL for G12
6	172	35	1.1	1.0	1.0	1.0
7	214	34	1.4	1.4	1.3	1.3
8	224	33	1.5	1.4	1.4	1.4
9	234	32	1.5	1.5	1.5	1.4
10	243	31	1.6	1.6	1.5	1.5
11	252	30	1.7	1.6	1.6	1.6
12	260	30	1.7	1.7	1.7	1.6
13	269	30	1.8	1.8	1.7	1.7
14	277	31	1.9	1.8	1.8	1.8
15	287	33	1.9	1.9	1.9	1.8
16	297	34	2.1	2.0	1.9	1.9
17	308	37	2.4	2.3	2.2	2.1
18	321	38	2.8	2.6	2.5	2.5
19	334	38	3.1	3.0	2.9	2.8
20	347	37	3.3	3.2	3.1	3.1
21	359	35	3.5	3.4	3.3	3.2
22	370	34	3.7	3.6	3.5	3.4
23	380	33	3.9	3.7	3.6	3.6
24	391	33	4.1	3.9	3.8	3.7
25	401	34	4.2	4.1	4.0	3.9
26	412	36	4.4	4.3	4.2	4.1
27	425	39	4.7	4.6	4.4	4.3
28	438	44	4.9	4.8	4.7	4.6
29	455	53	5.5	5.3	5.1	5.0
30	476	72	6.0	6.0	6.0	6.0

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

2.7 Equating Summary

Each year a certain percentage of items on each ACCESS for ELLs test form are new, as determined by the refreshment plan for that series. For Series 501, we refreshed all four domains. Many items appearing on ACCESS Online Series 501 also appeared on Series 403. These items are referred to as common items. The number of common items between series by domain depends on the targeted refreshment plan for the particular series and domain. We use an equating procedure known as common-item equating to equate the results on new forms to the older forms using the common items. In this procedure, we keep the difficulty measures for items that appear on both the new and the old forms constant across both forms. Thus, performances on the newer form may be interpreted using the same frame of reference. We anchored all items common to both forms to their 403 values in the first equating run except for the Writing domain. Series 501 saw a redesign of the ACCESS Writing test, in which the number of Writing tasks was reduced from either three or four per form to two per form. The parameters of the Writing domain Series 403 continuing tasks were anchored to their values derived on a two-task scale in a special research study (CAL, 2019). For the Speaking domain, we also anchored difficulty measures for the new tasks to their initial calibrated values from the Speaking field test analysis.

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 level 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 it not been anchored. 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, 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 Speaking tasks, a slightly different displacement criterion (above 0.50 or below -0.50) was used since anchored tasks from the Speaking domain have been shown to be less stable than items and tasks from other domains.

A pre-equating design was used to conduct the annual equating for Listening and Reading. This design allows 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, student data collected from the Series 501 embedded field test were used to conduct the equating analyses. All available student data at the time the equating analyses were conducted were included in the analyses.

For the Writing domain, the annual equating analysis was conducted using 501 operational data collected during the early testing window. The Writing equating study was conducted with a random sample drawn with a target sample size of 1,500 for Tier A forms, 1,500 for Grades 1–5 Tier B/C forms, and 2,000 for Grades 6–12 Tier B/C forms. The Writing equating sample was

drawn so that it was proportional to the Series 402 and Series 403 population means for the Writing domain, by grade and tiered form.

For the Speaking domain, student data from the Series 501 appended Speaking field test administration were used to conduct the initial common-item equating. These initial item parameters were then verified using Series 501 operational data collected during the early testing window. The Speaking verification study was conducted using a random sample drawn with a target sample size of 3,000 students per grade-level cluster. The Speaking verification sample was drawn such that it was proportional to the Series 401, 402, and 403 population means for the Speaking domain, by grade and tier.

Tables in this section present a summary of the equating and verification procedures. The first section of the tables compares the current test (i.e., the Series 501 version of that test form) to the previous year's test (i.e., the Series 403 version of that test form). 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 are shown. These values are in log-odd units, or "logits" (i.e., analyses carried out using the Rasch measurement model, which produce equal-interval, linear measures expressed on a logit scale). In the domains of Listening and Reading, if the equating was successful, we expect the average item difficulty values between 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. In Writing, we expect some differences in the average difficulty values between series. As mentioned above, there was a Writing test redesign in Series 501, and tasks were anchored not to the Writing Series 403 operational values, but to values derived in a special study to ensure a smooth transition to the new test design. Because of this we expect some difference between Series 501 and the operational 403 values. Additionally, the 501 Writing domain tests consist of only two tasks, with only one task serving as an anchor between series, with the exception of grade-level cluster 6-8 Tier B/C, which has two anchor tasks. Similarly, we might expect some differences in average difficulty values between series for Speaking, as test forms consist of only nine tasks, and only one-third of the test serves as the anchor between series.

The second section of the tables presents information on the anchoring items. The total number of possible anchors that were initially anchored to the value of the previous series is shown, as well as the average item difficulty and the average standard deviation of the difficulty values for those items. Next, the number of items that were anchored in the final equating run is shown, again with the average item difficulty and the average standard deviation of those difficulty values for those items. Finally, the percentage of items that served as anchors and their average displacement values are given. In general, the larger the number and the higher the percentage of items anchored and the closer their average displacement is to 0.00, the more trustworthy the equating results will be. For the Listening and Reading domains, the average displacements are expected to be around 0.00 since there are high percentages of items anchored. For the Writing domain, when there is only one task anchored to the known value derived from the special

research study, the displacement statistic for the anchor task is automatically set to 0 in Winsteps and the average displacement statistic is also 0.

The third section of the tables gives information about the anchor items, both by order of displacement statistics and by order of item difficulty. The displacement statistics provide information on the difference between the difficulty value of the anchored item and what that difficulty value would have been had the item not been anchored. Smaller displacement statistics indicate more consistency between the item's difficulty value on the Series 501 test form and on the Series 403 test form.

For longer tests such as Listening and Reading, it is desirable that the anchor items represent a wide range of difficulties across the entire spectrum of the item difficulty values on a test form. In addition, the greater the representation across the difficulty range of anchor items, the more trustworthy the equating results will be.

For the Writing and Speaking domains, which are shorter and performance based, and which have additional content and exposure considerations in terms of item refreshment, this rule of thumb may not always apply. In addition, the number of anchors is also a function of the targeted refreshment plan, which can differ by series and by domains.

For the Writing and Speaking tasks, this table has a fourth section, which provides the anchored Rasch rating scale model step measures for each task. For the ACCESS Writing and Speaking tasks, a Rasch-grouped rating scale model is used (see the introduction to Section 2). The step measure corresponds to the location on the latent variable where the probability of students receiving a rating category and the category below it on the rating scale are equal, relative to the difficulty measure of the task. Step measures indicate how likely it is to observe a category relative to other categories on the scale and do not indicate the difficulty measure of the category (Linacre, 2004). The step measures are expected to increase with category values. If the step measures do not increase in value up the rating scale, then it indicates that the frequency of the category is small relative to those of other categories since the category only occupies a narrow range on the latent scale. For Writing, there is only one rating scale being modeled and step measures are the same for all the tasks. Writing step measures increase up the rating scale except for score point 3, which indicates a lower frequency of occurrence for score point 3 compared to other categories. To provide anchors in the calibration of new Writing tasks to facilitate their placement onto the common WIDA score scale each year, the same step measures are held constant. For Speaking, all PL 1 tasks are modeled by one rating scale and all PL 3 and 5 tasks are modeled by a different rating scale. Therefore, the step measures for all PL 1 tasks are the same, and the step measures for all PL 3 and PL 5 tasks are the same. In Speaking, all step measures increase with category values for both rating scales. 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.

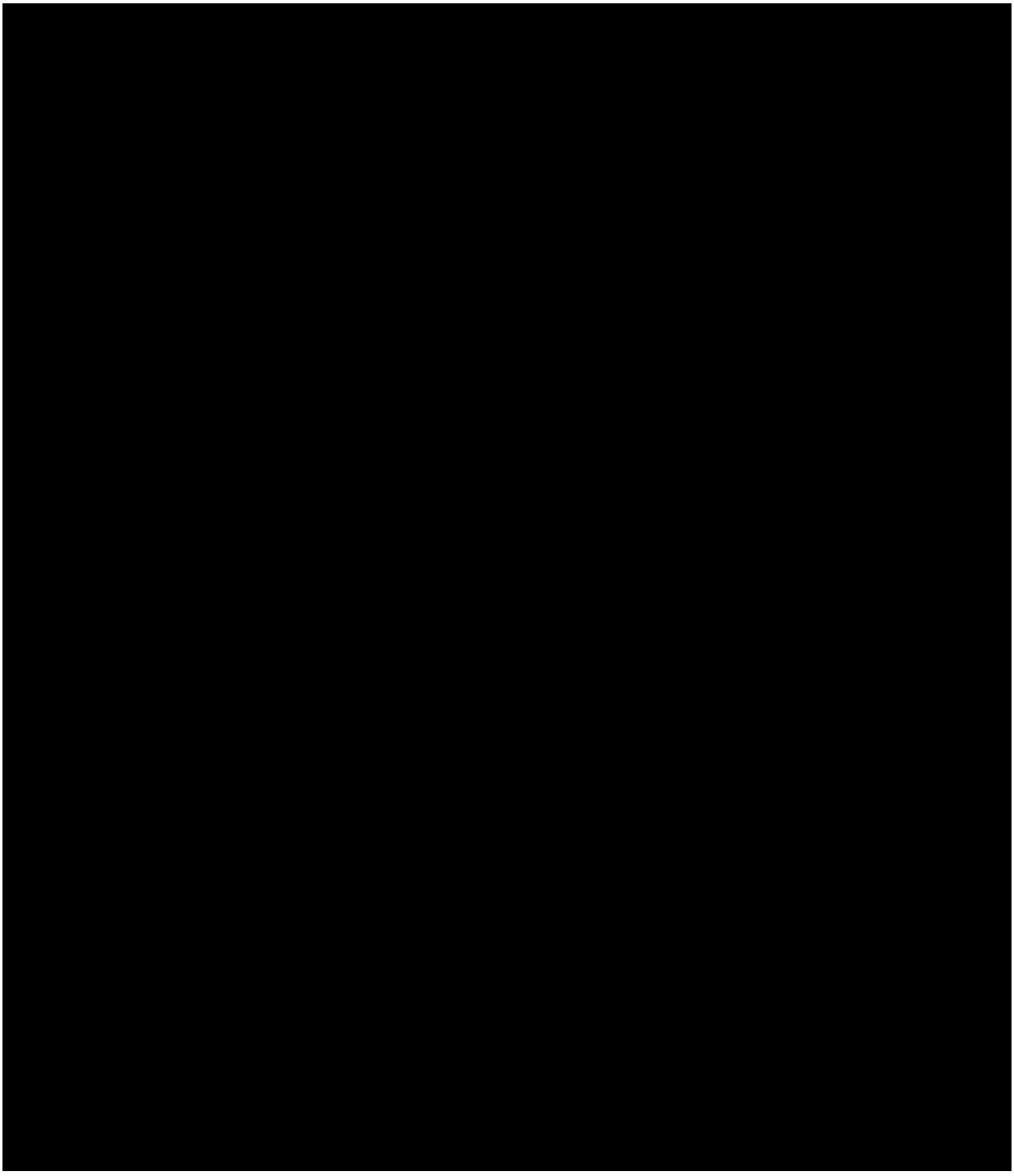
The results show that the average difficulty levels of the 501 Listening and Reading item pools are similar to those of the previous series for all grade -clusters, except for Reading grade-level

cluster 2-3. This was due to the 501 strategic refreshment plan, which called for a slight increase in the difficulty level that the item pool was targeting. The average difficulty levels of the Writing domain are similar to those of the previous series for all grade clusters and tiers, except for Grade 1 Tier A, Grade 1 Tier B/C, and Grades 2–3 Tier B/C. This is because the number of tasks was reduced from four or three to two, and the easiest items of these forms have been removed from 501. The average difficulty levels of the 501 Speaking domain are similar to those of the previous series for all grade clusters.

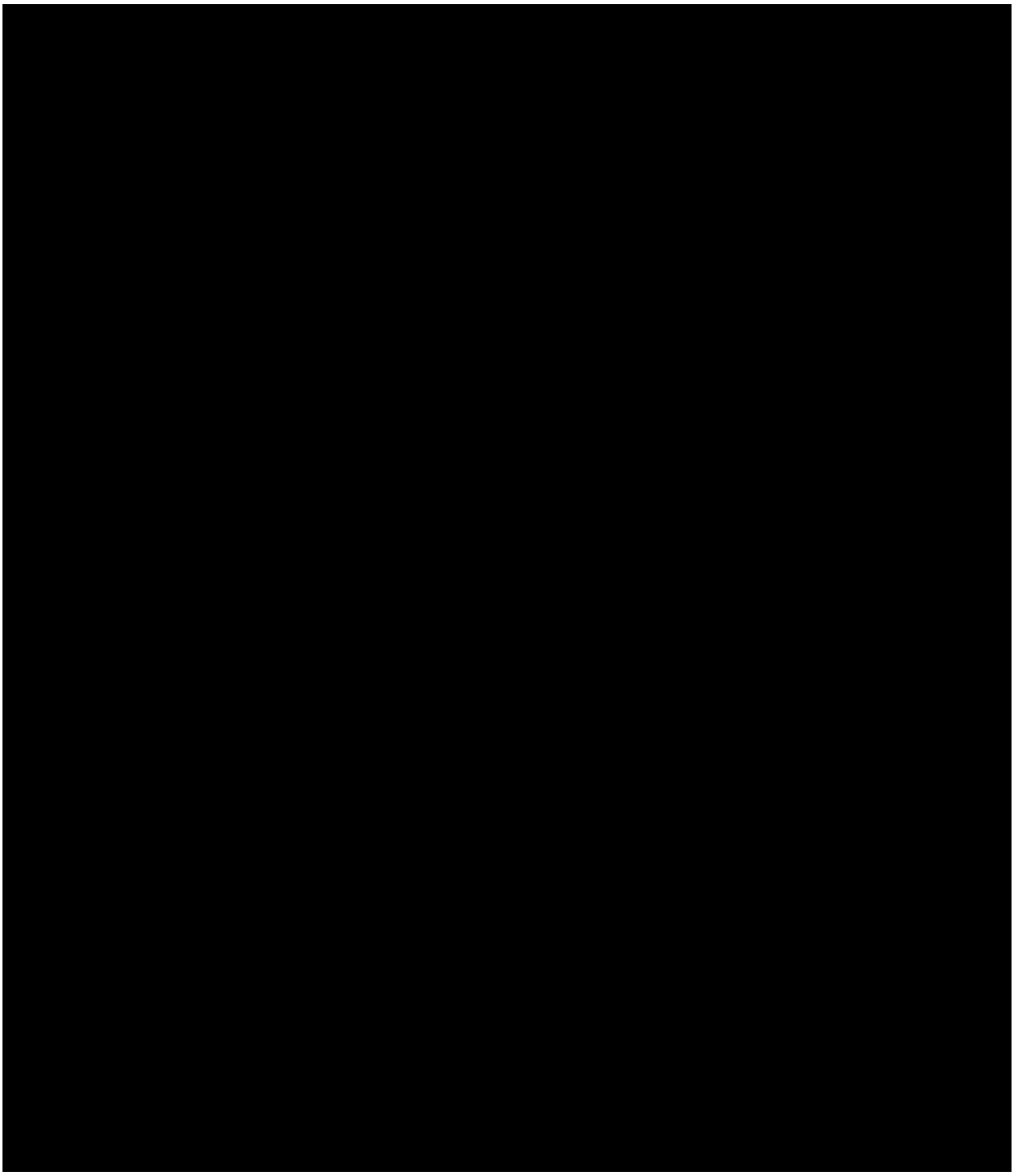
For the Listening domain, the percentage of items anchored in the final equating run ranged from 61% to 76% and the average displacement statistics were either equal to or close to 0.00. For the Reading domain, the percentage of items anchored in the final equating run ranged from 60% to 72% and the average displacement statistics were either equal to or close to 0.00. For the Writing domain, all test forms except for grade-level cluster 6–8 Tier B/C had only one task anchored to the known value derived from the special research study. The displacement statistic for the anchor task was automatically set to 0 in Winsteps and therefore the average displacement statistic is also 0. The average displacement statistics for Writing Grades 6–8 Tier B/C is -0.03. For the Speaking domain, the percentage of tasks anchored in the final equating run was between 33% and 67% and the average displacement statistics were all close to 0.00.

2.7.1 **Listening**

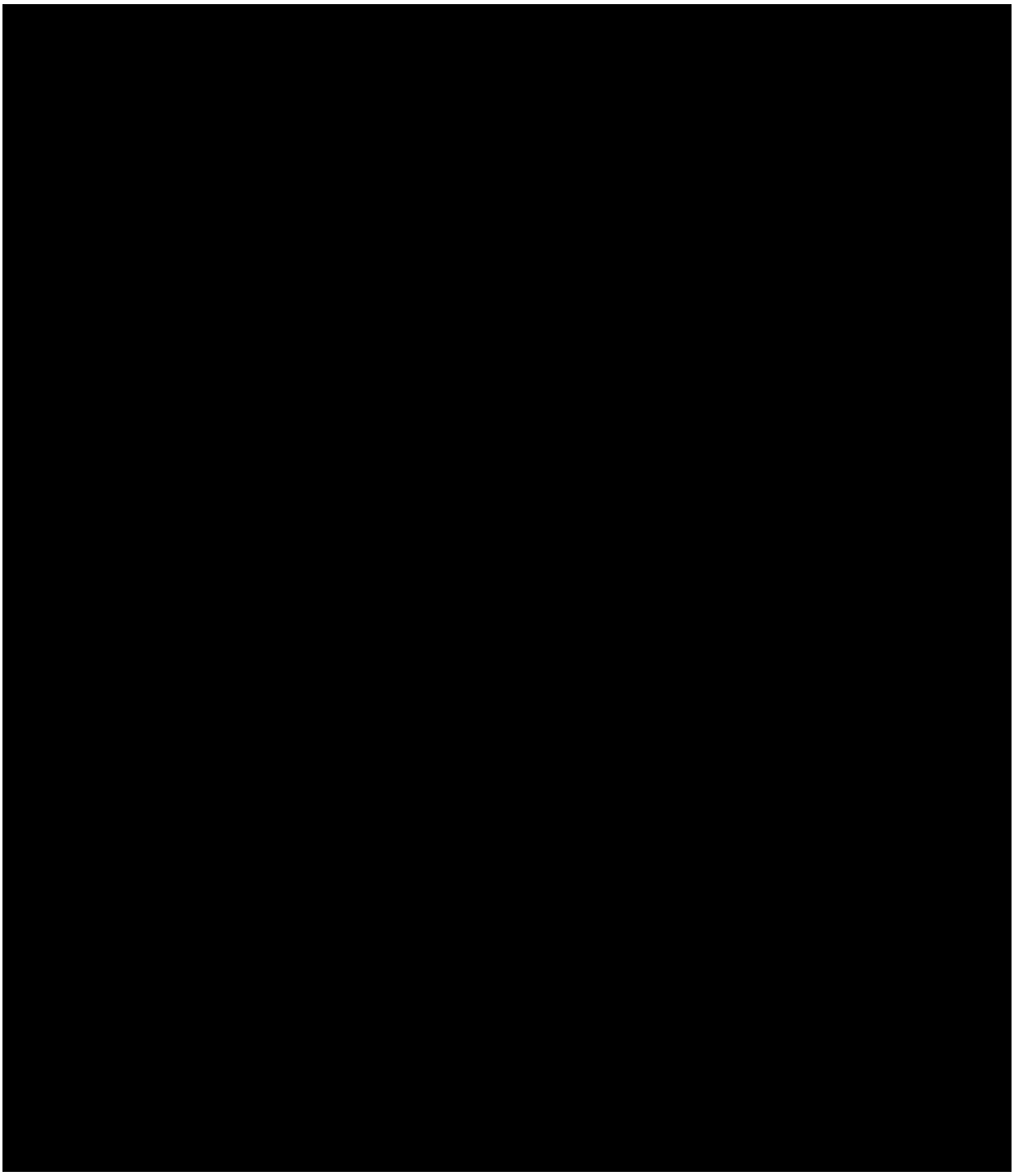
2.7.1.1 *Grade 1*



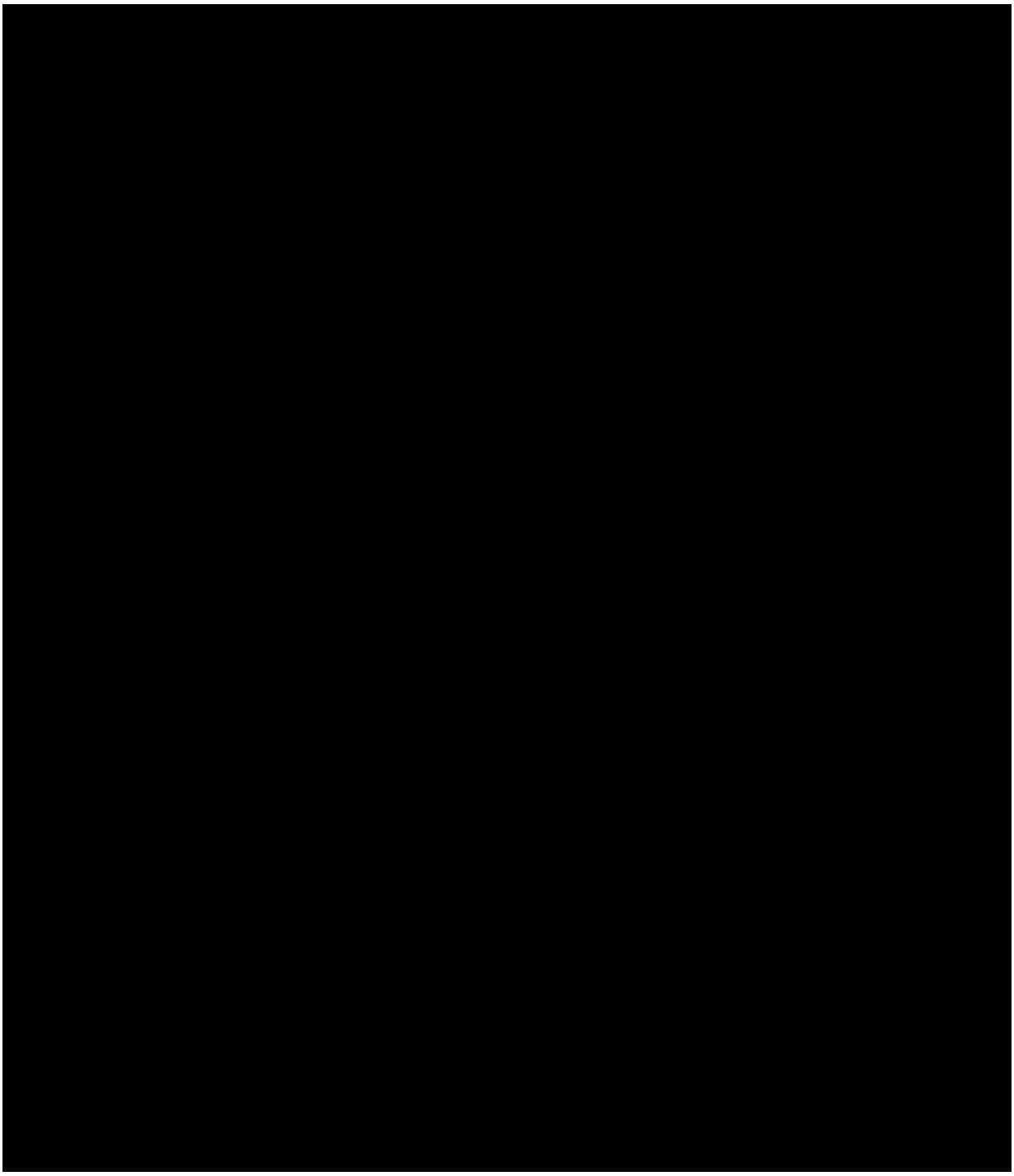
2.7.1.2 *Grades 2–3*



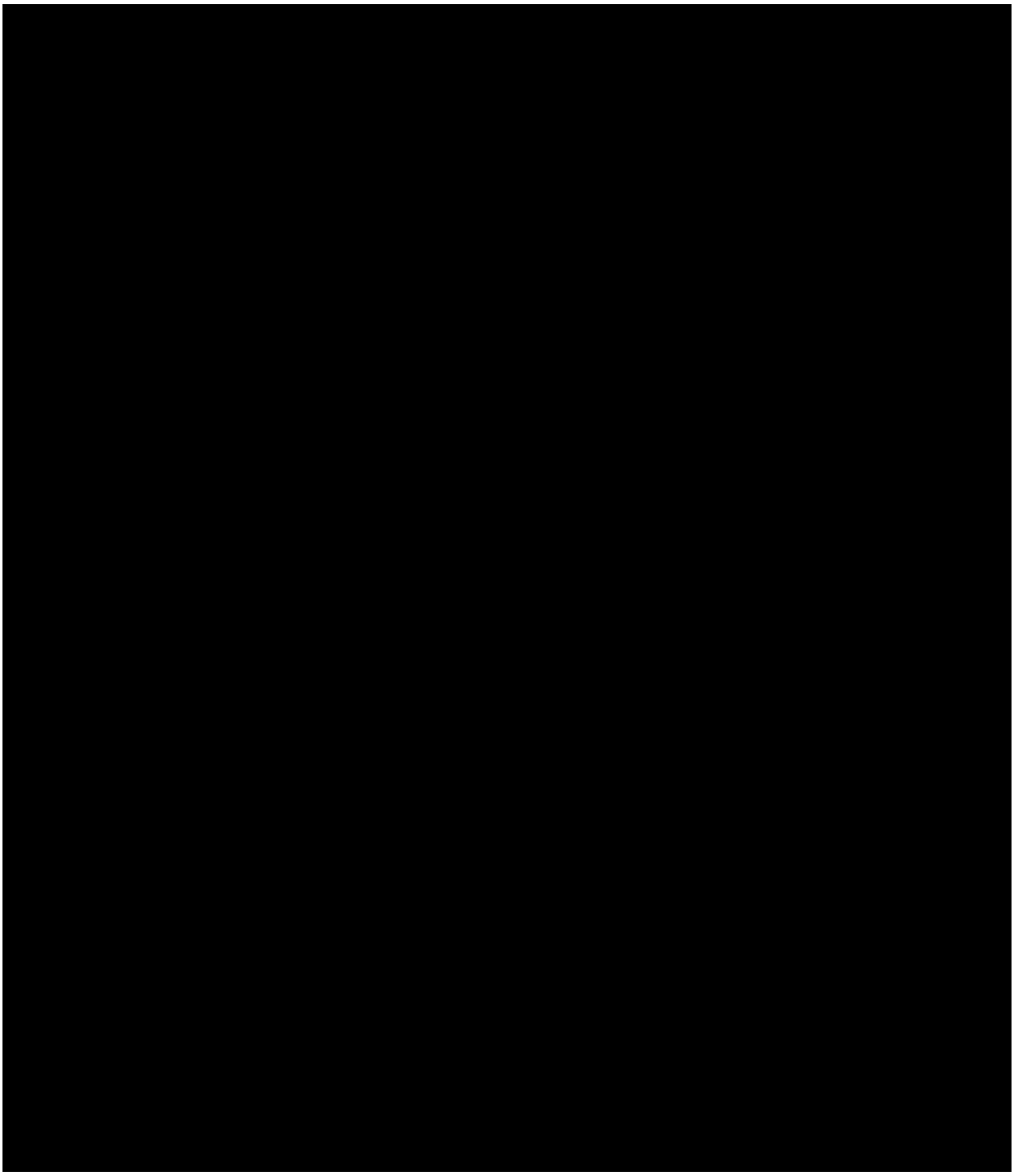
2.7.1.3 *Grades 4–5*



2.7.1.4 *Grades 6–8*

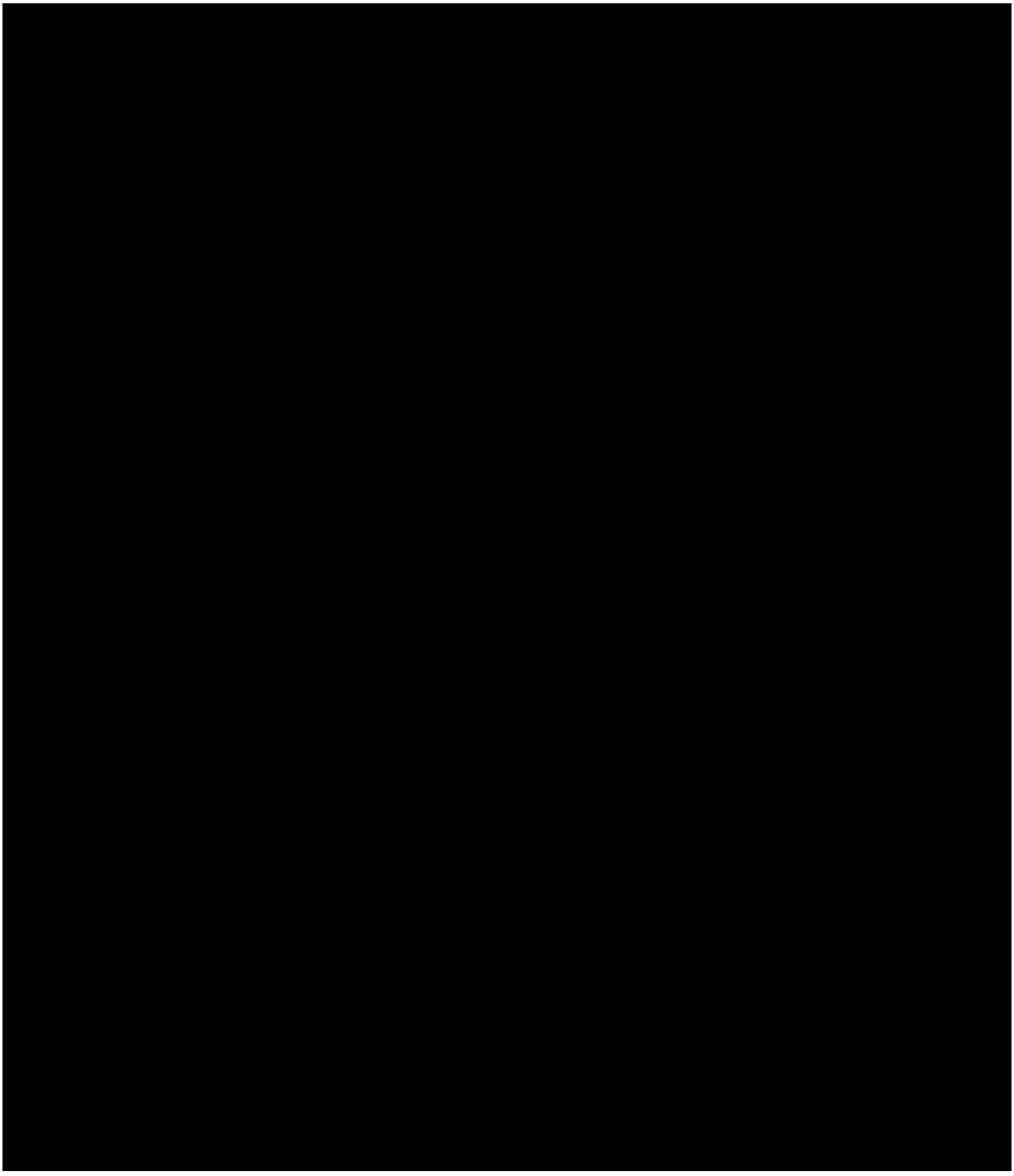


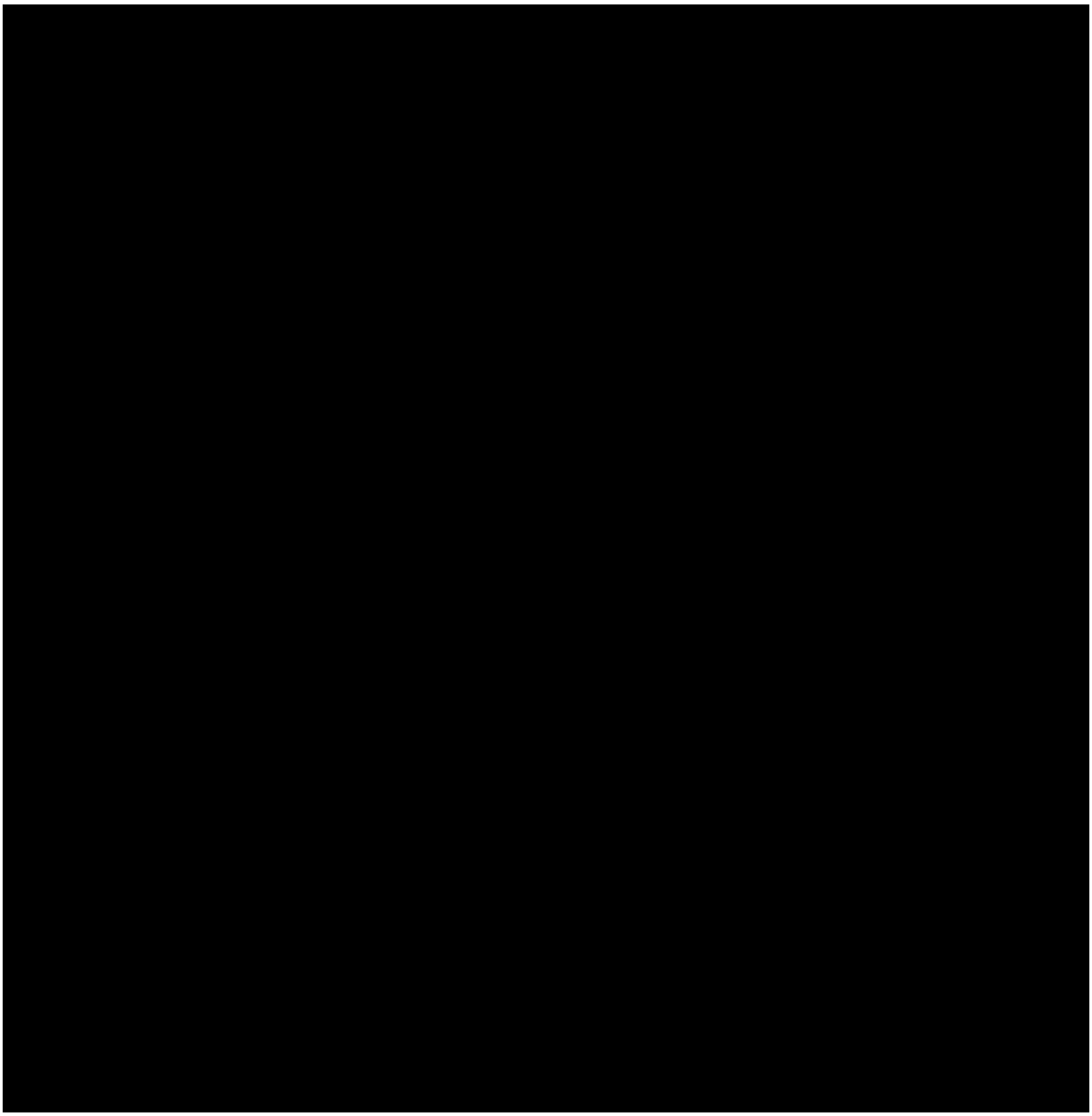
2.7.1.5 *Grades 9–12*



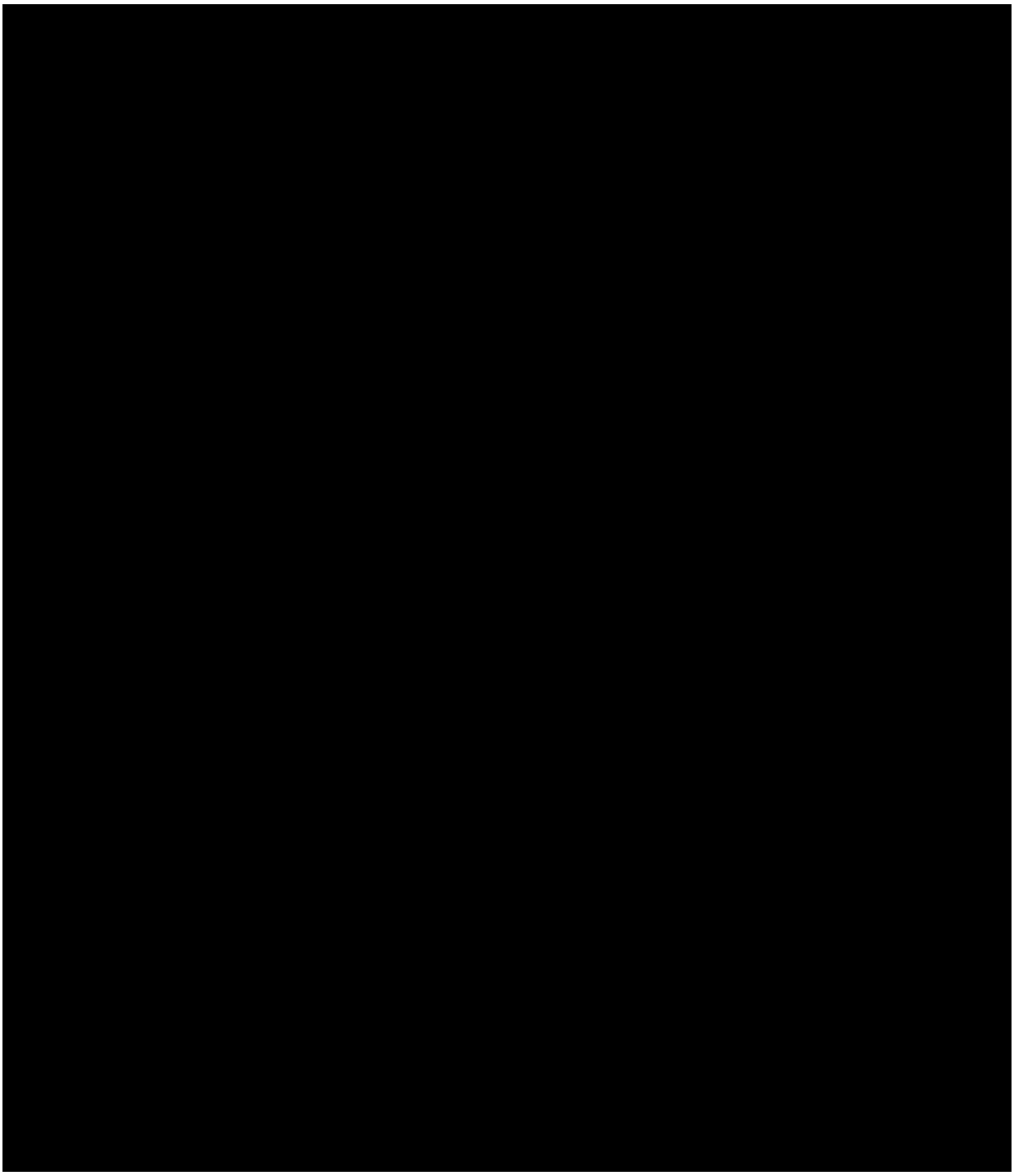
2.7.2 **Reading**

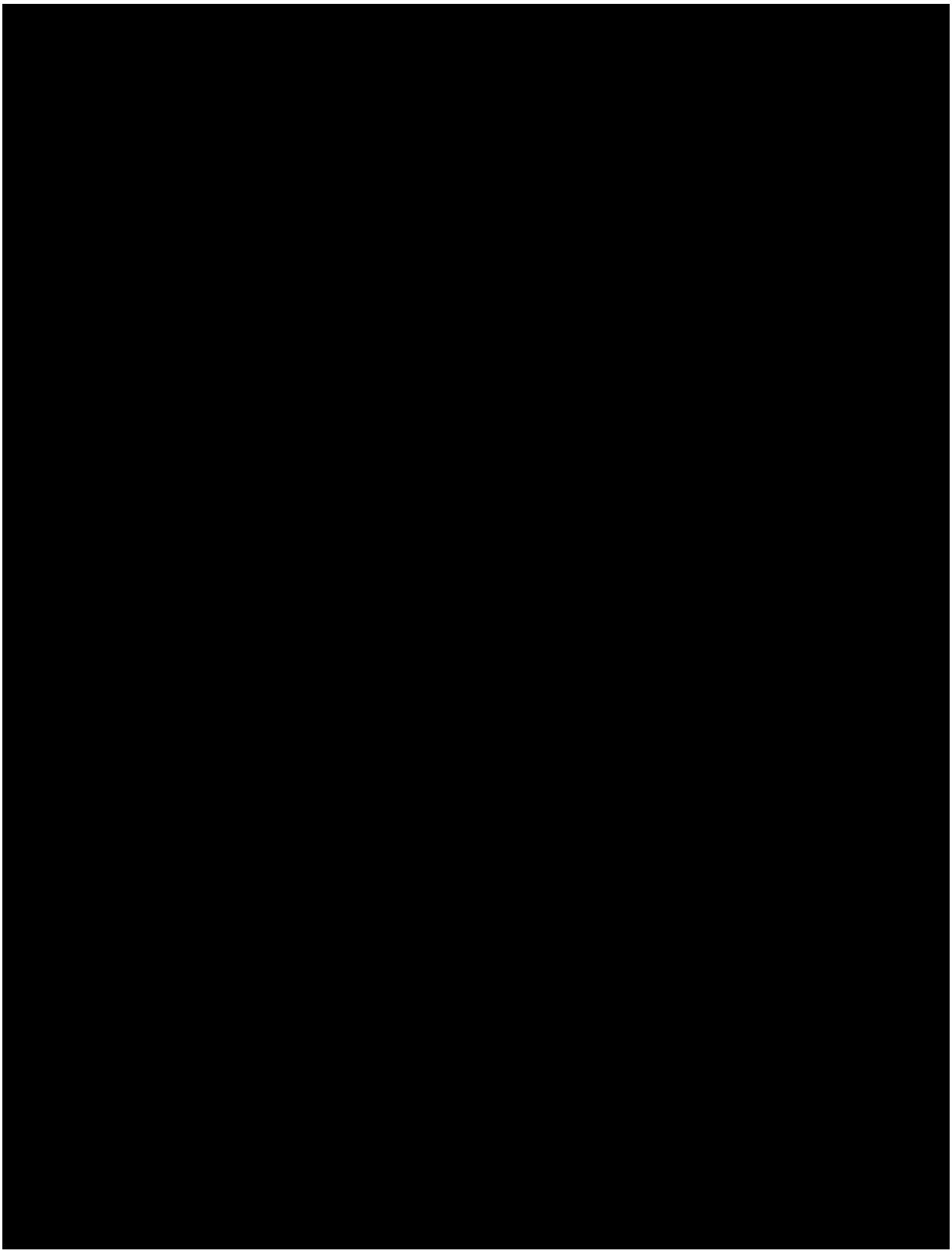
2.7.2.1 *Grade 1*



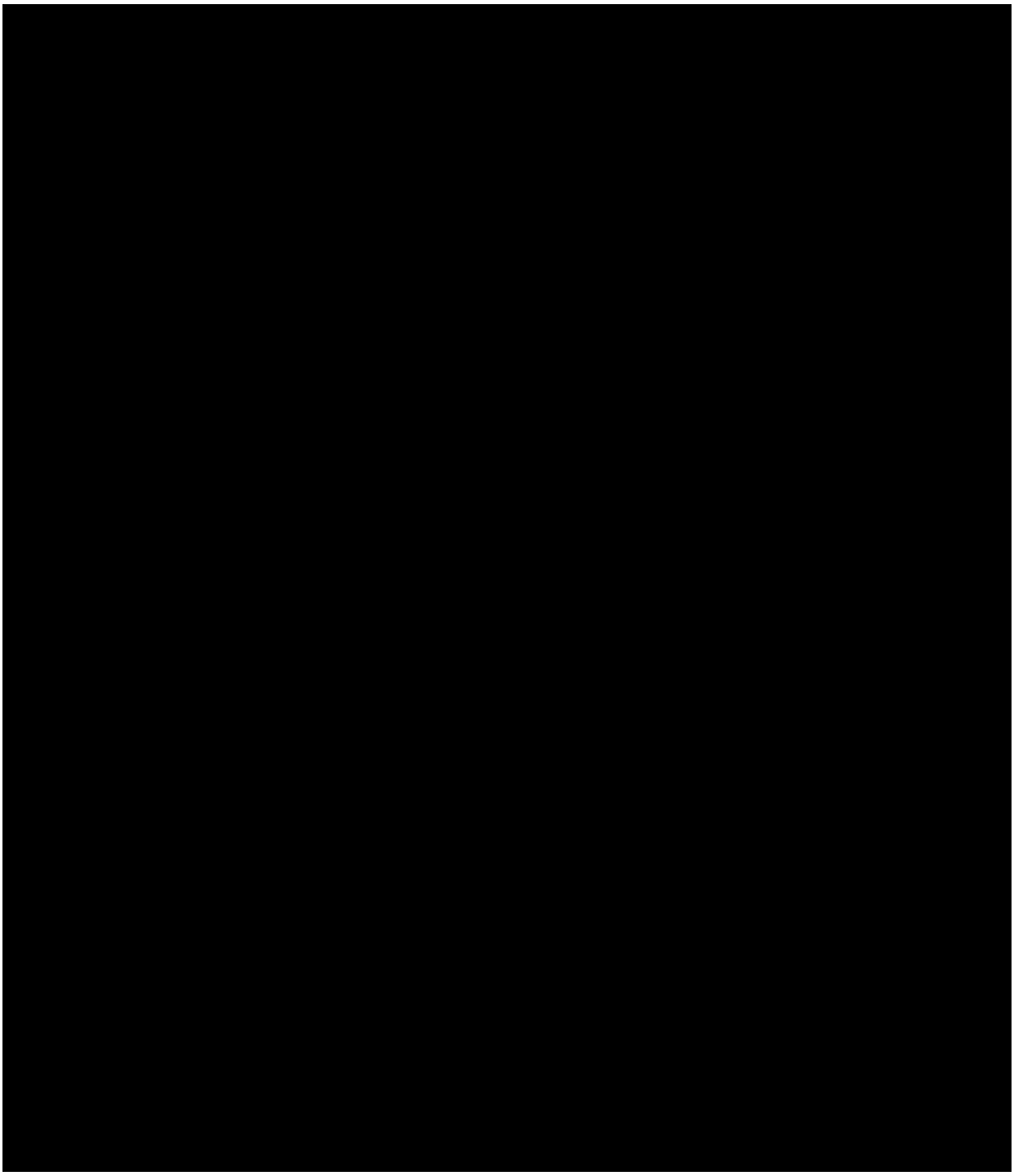


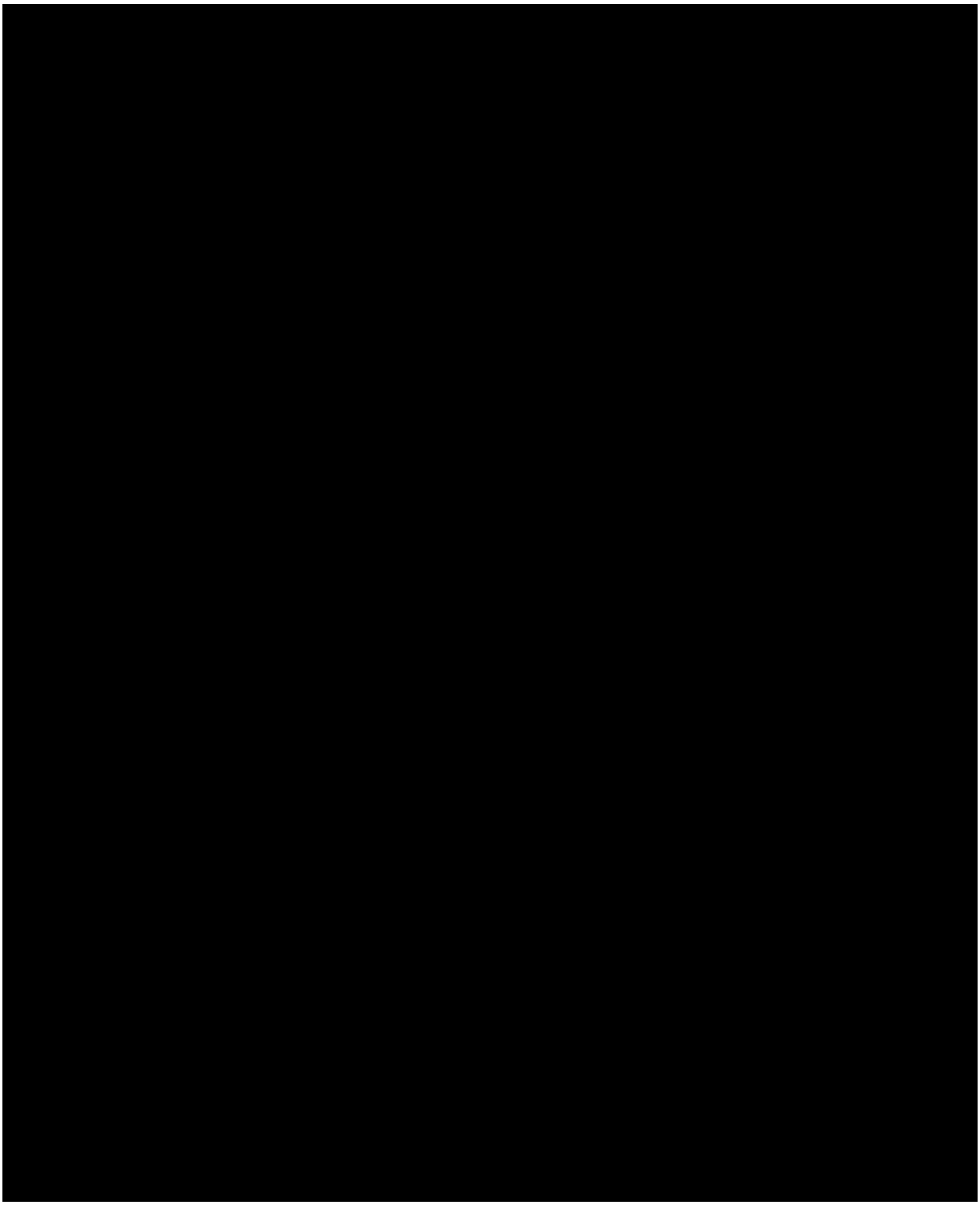
2.7.2.2 *Grades 2–3*



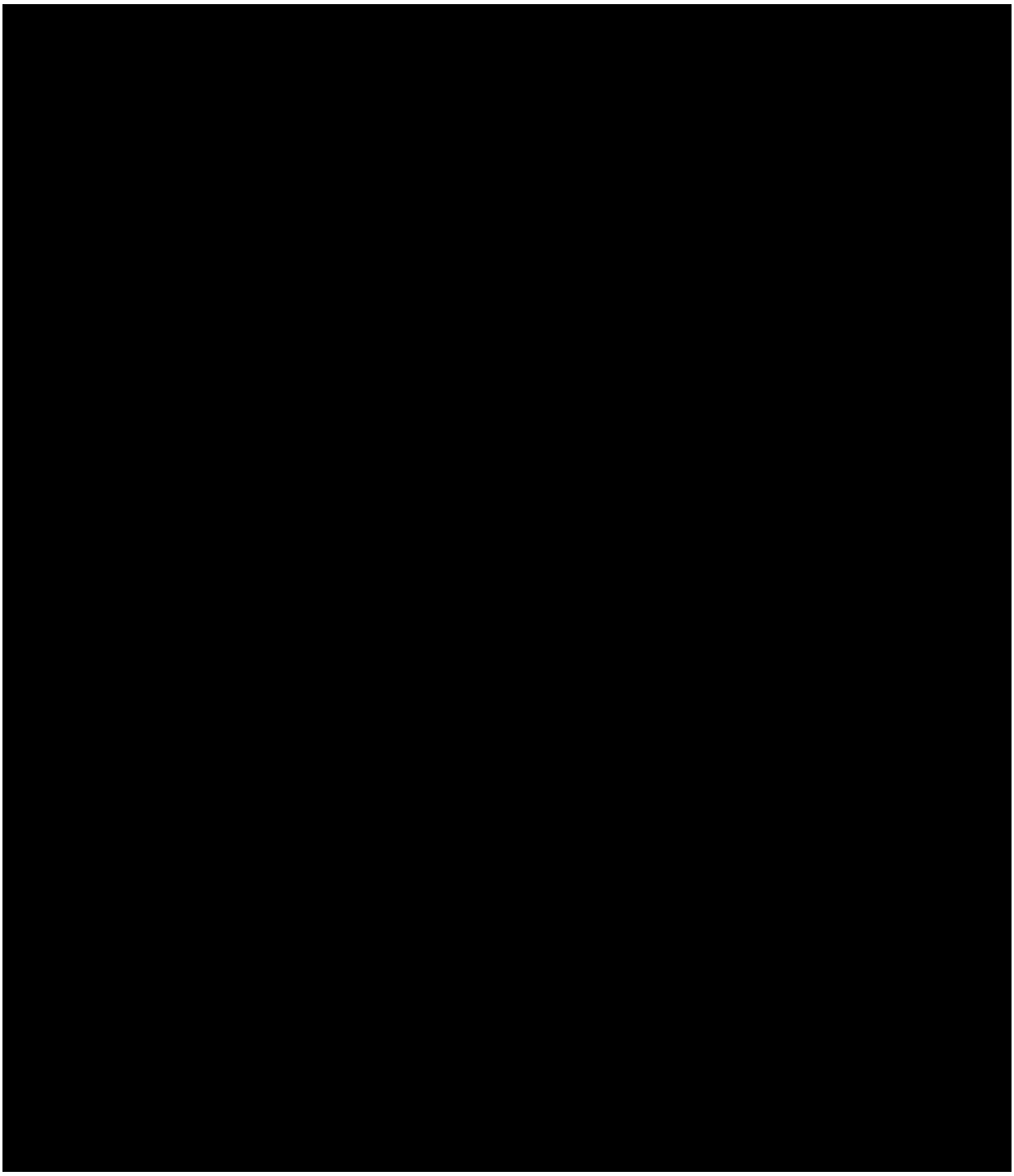


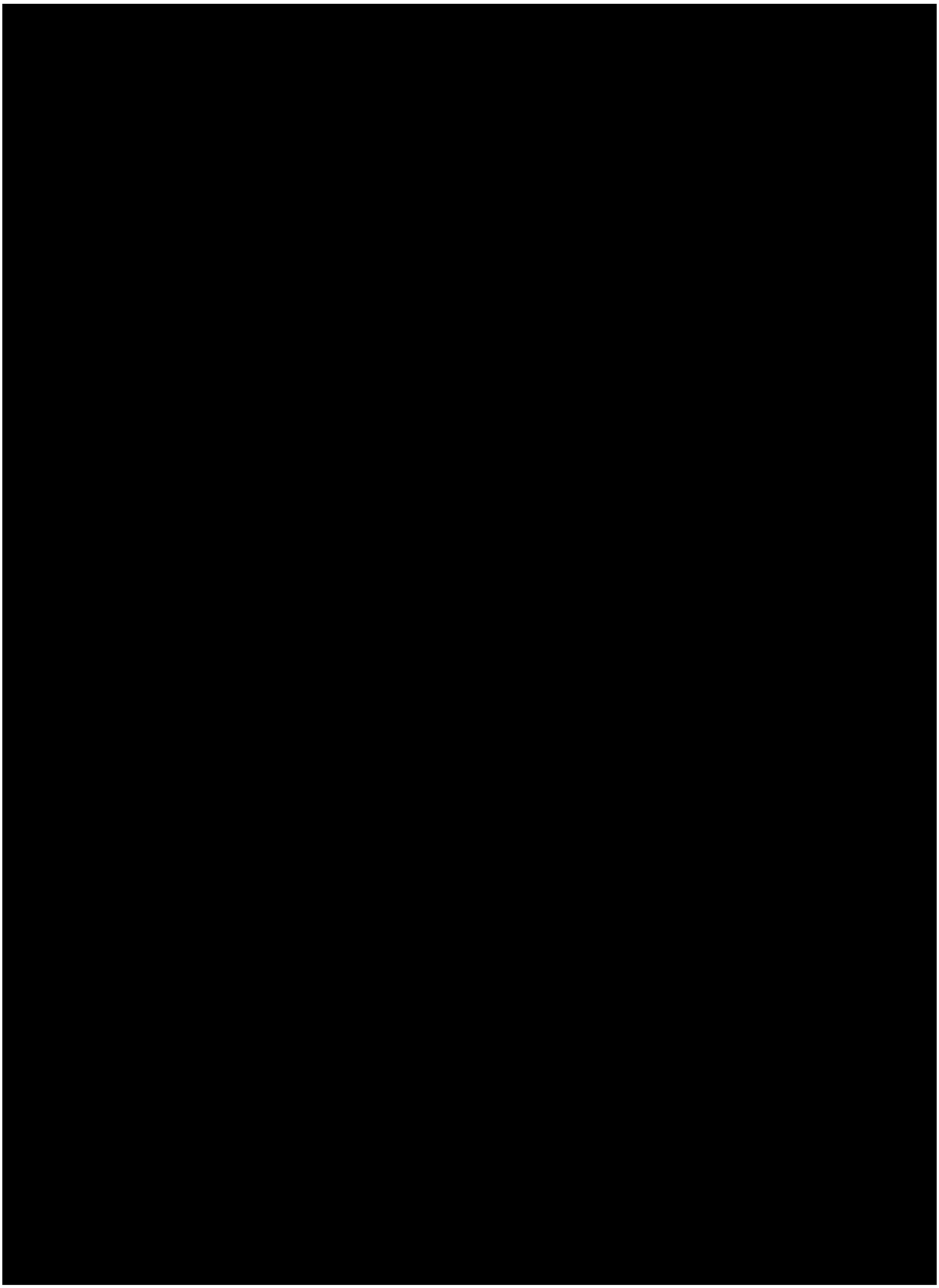
2.7.2.3 *Grades 4–5*



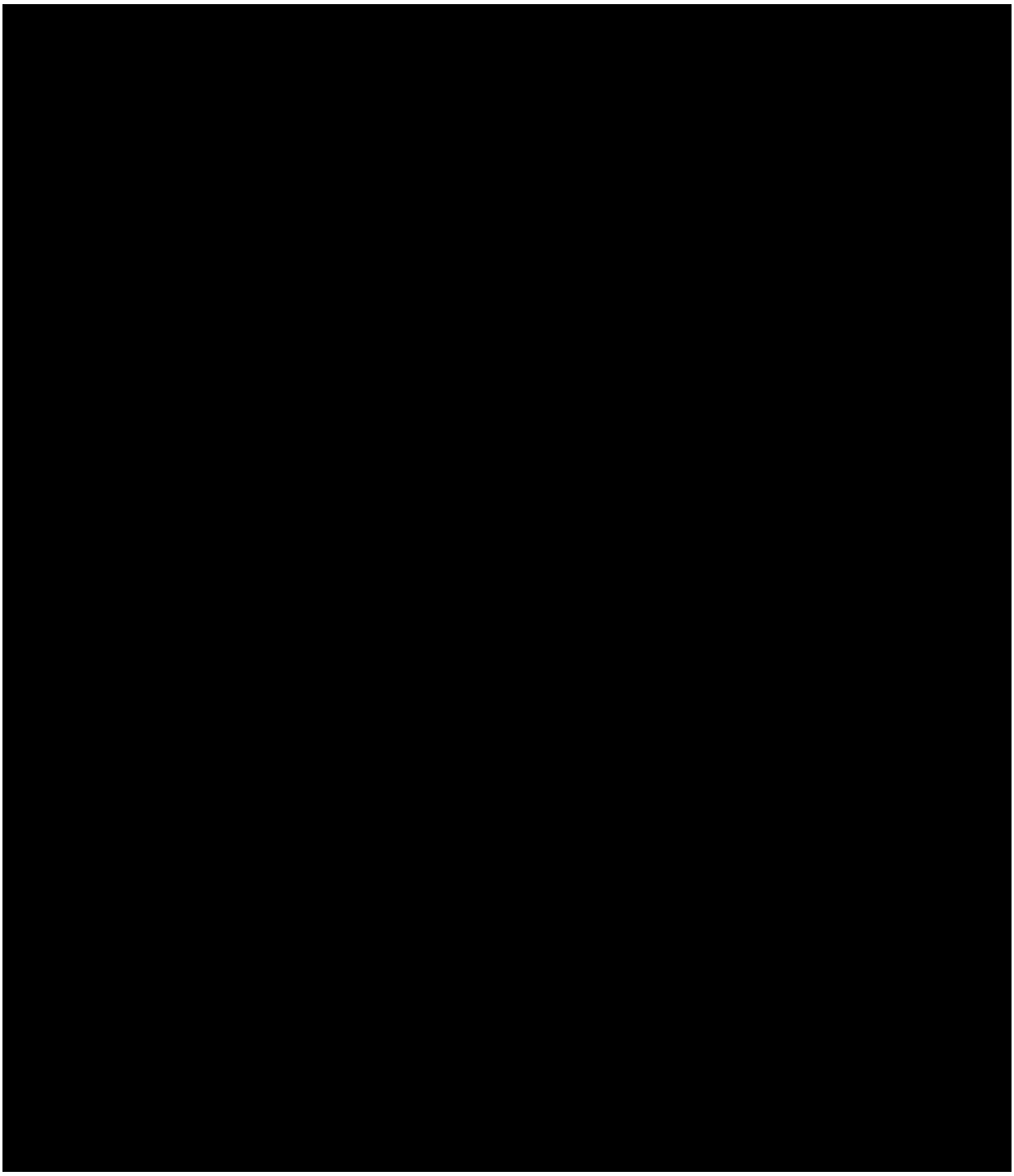


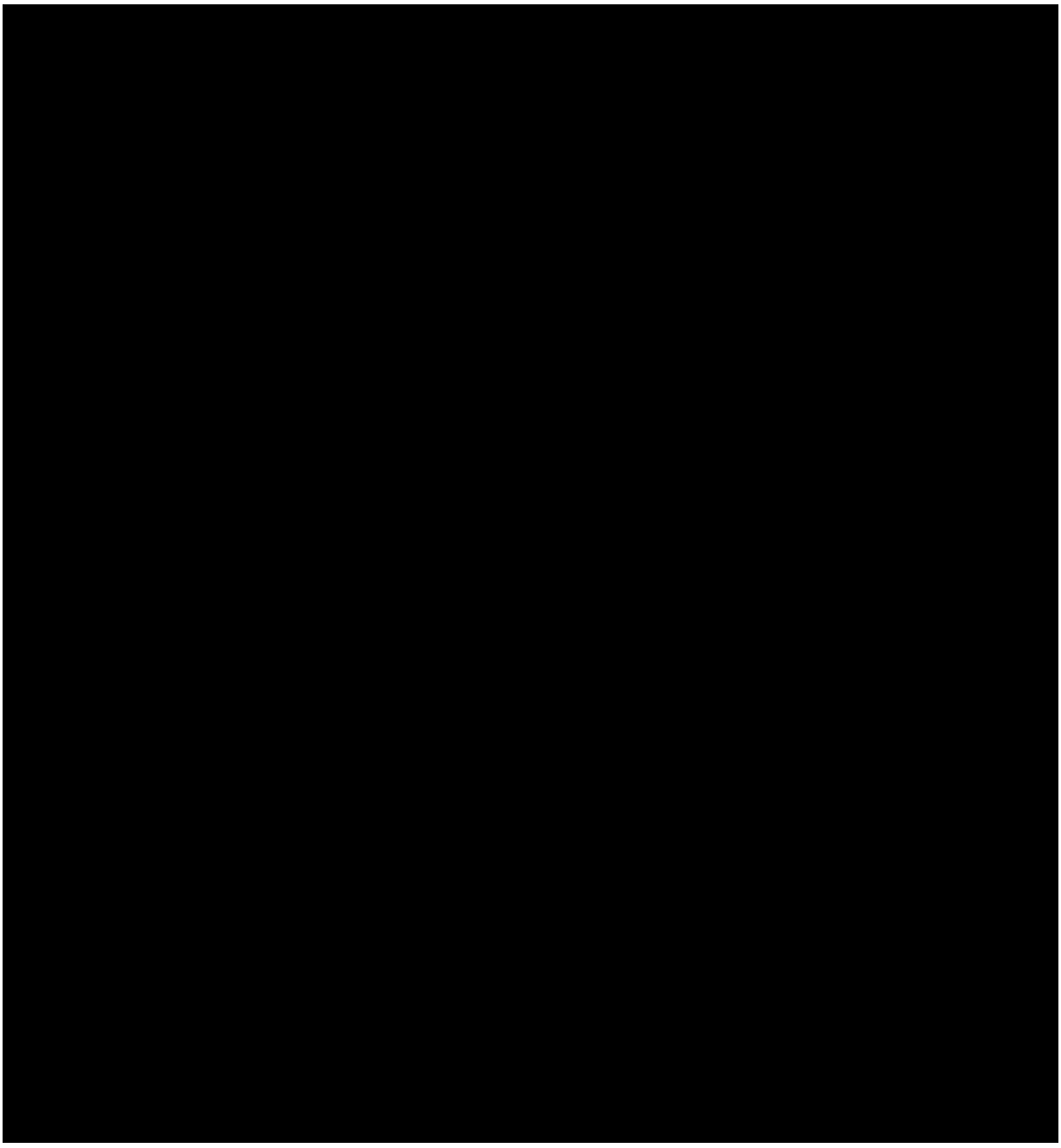
2.7.2.4 *Grades 6–8*





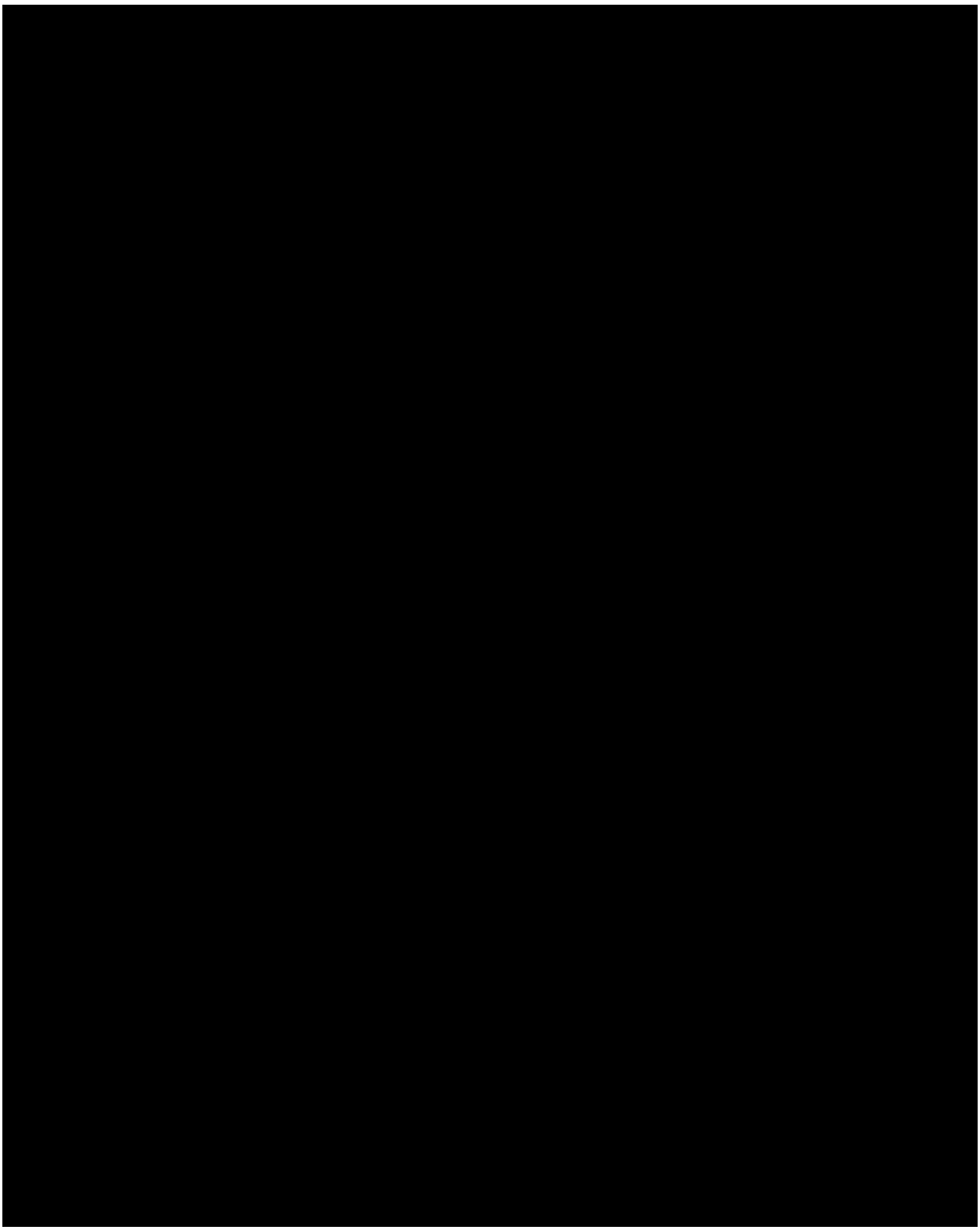
2.7.2.5 *Grades 9–12*

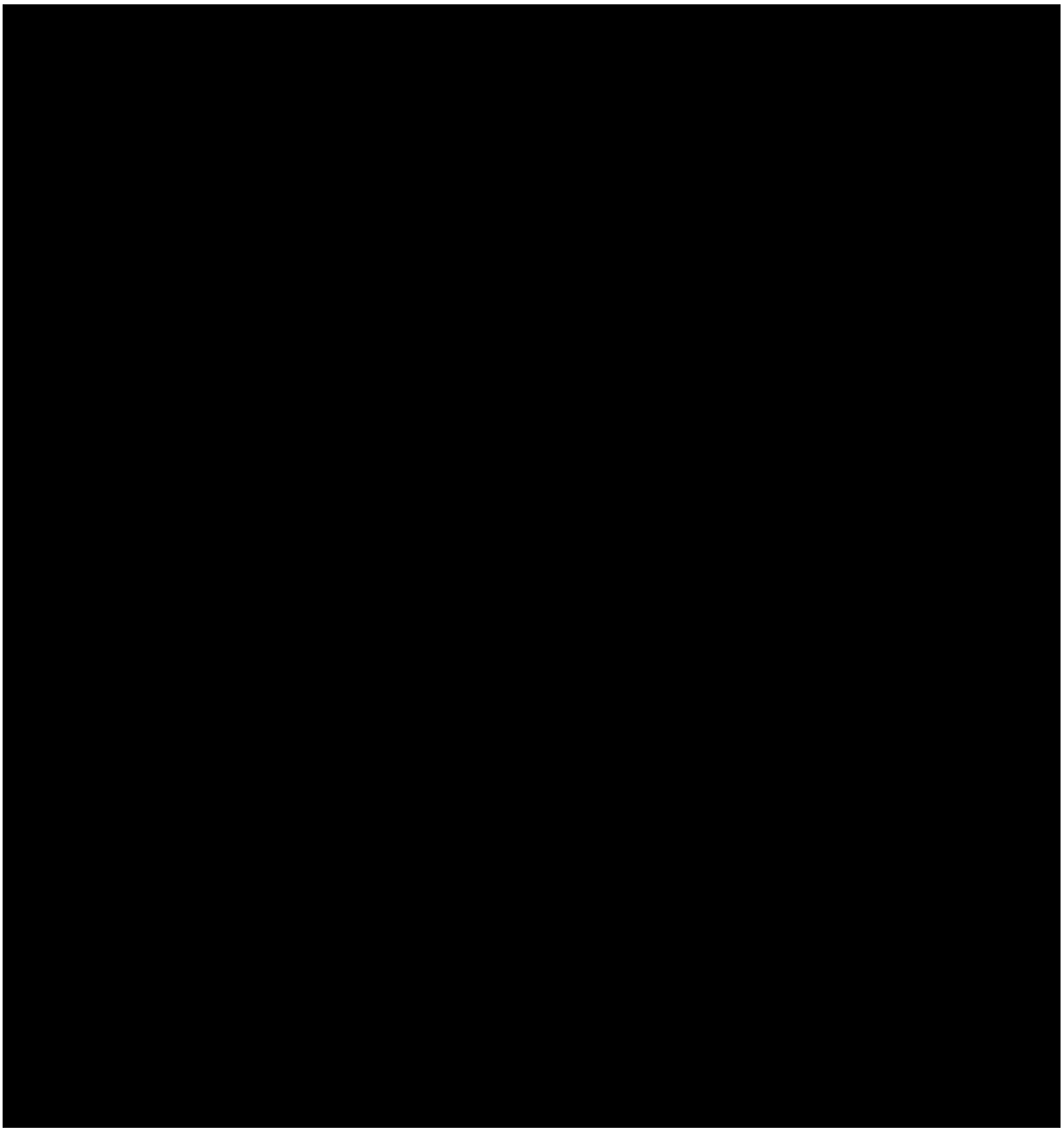




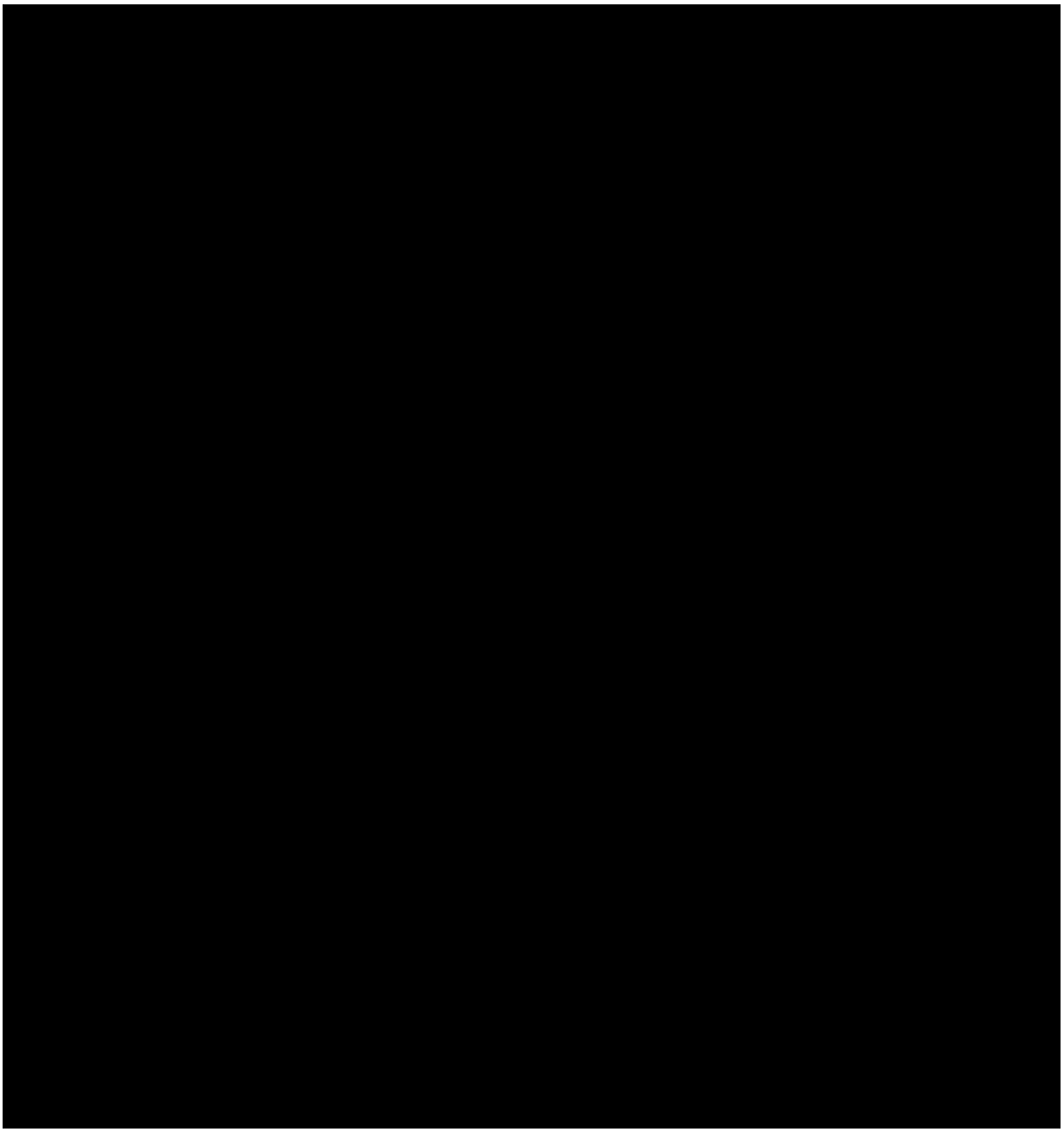
2.7.3 **Writing**

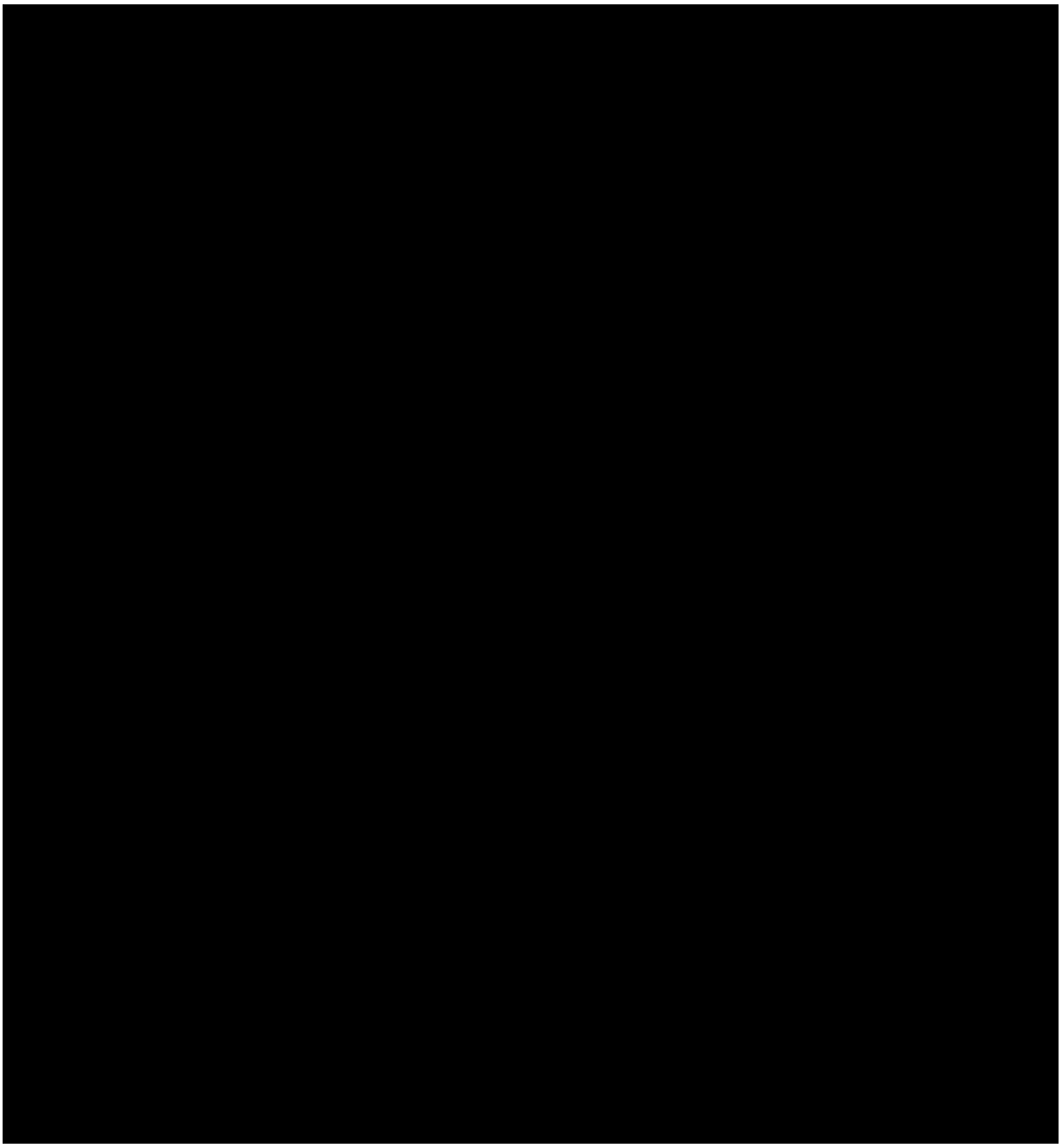
2.7.3.1 *Grade 1*



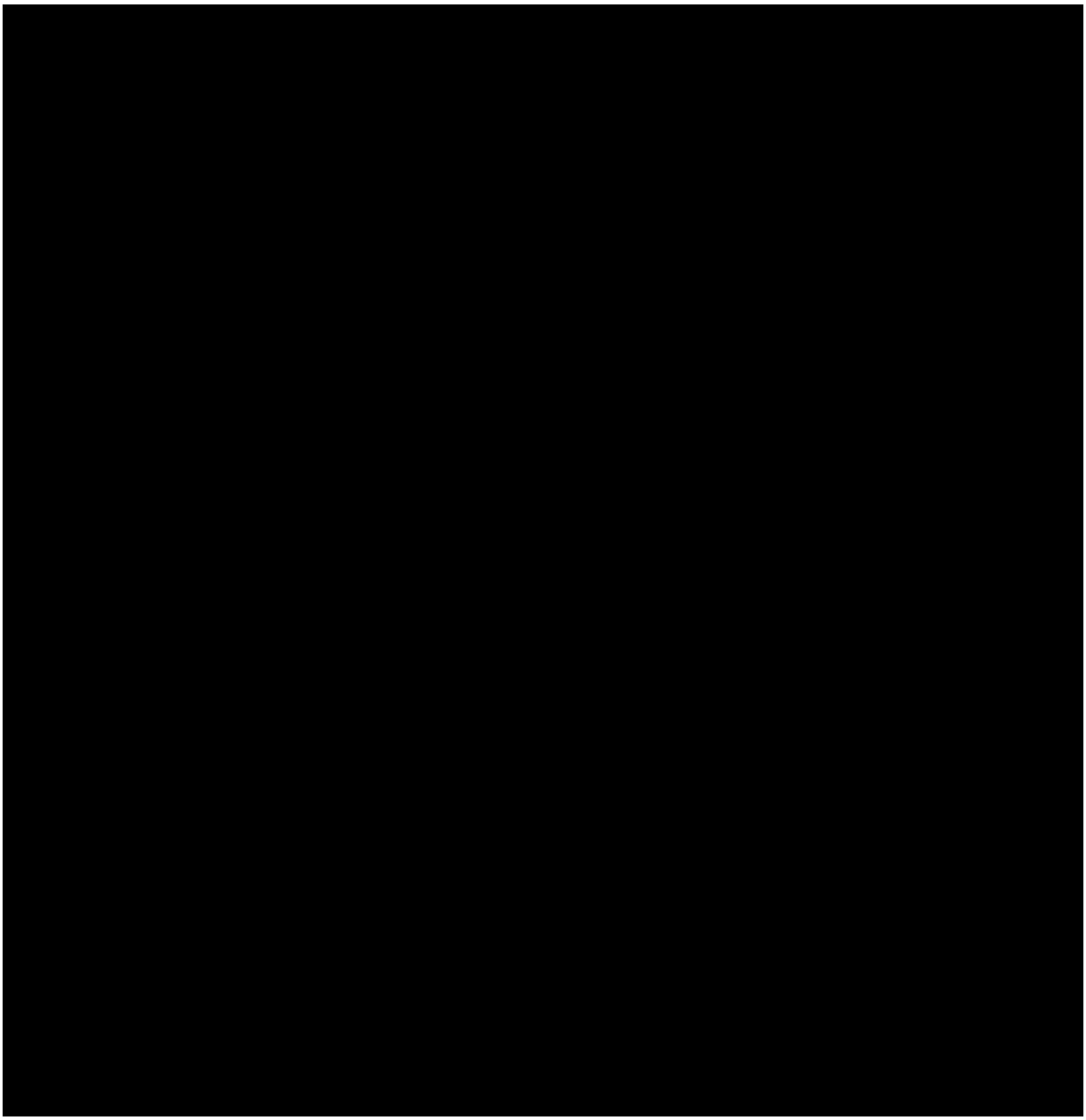


2.7.3.2 *Grades 2–3*





2.7.3.3 *Grades 4–5*



2.7.3.4 *Grades 6–8*

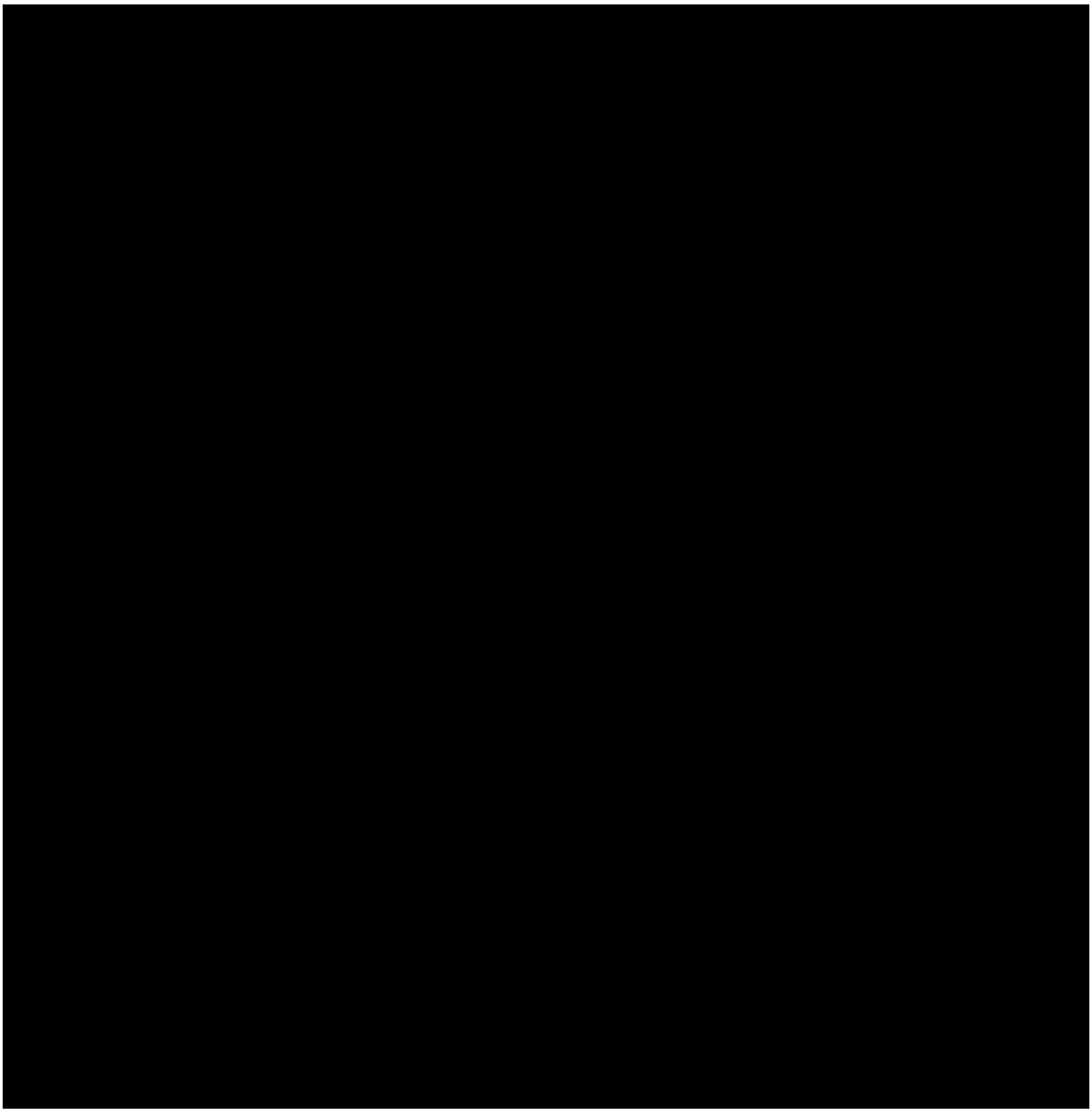
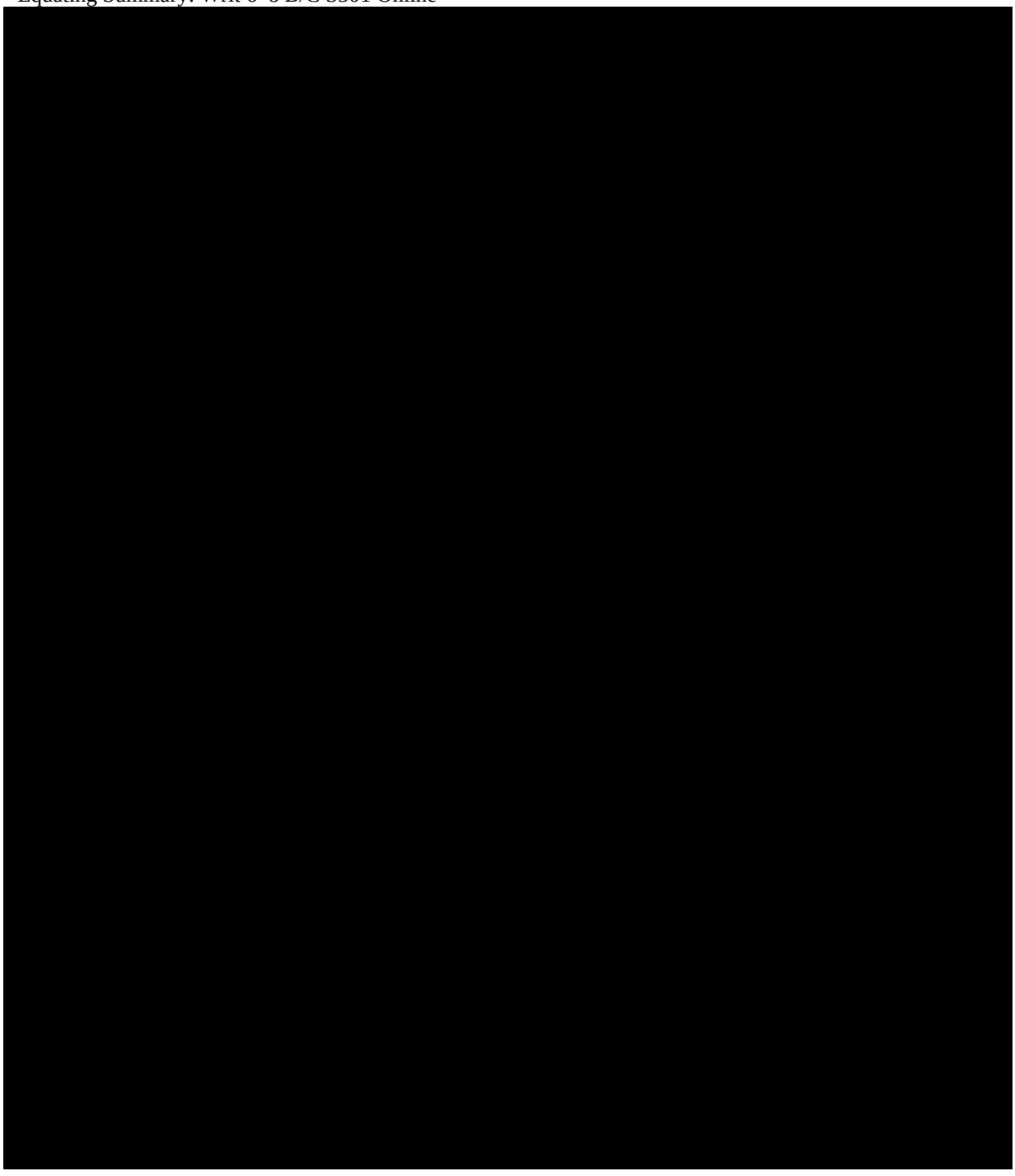
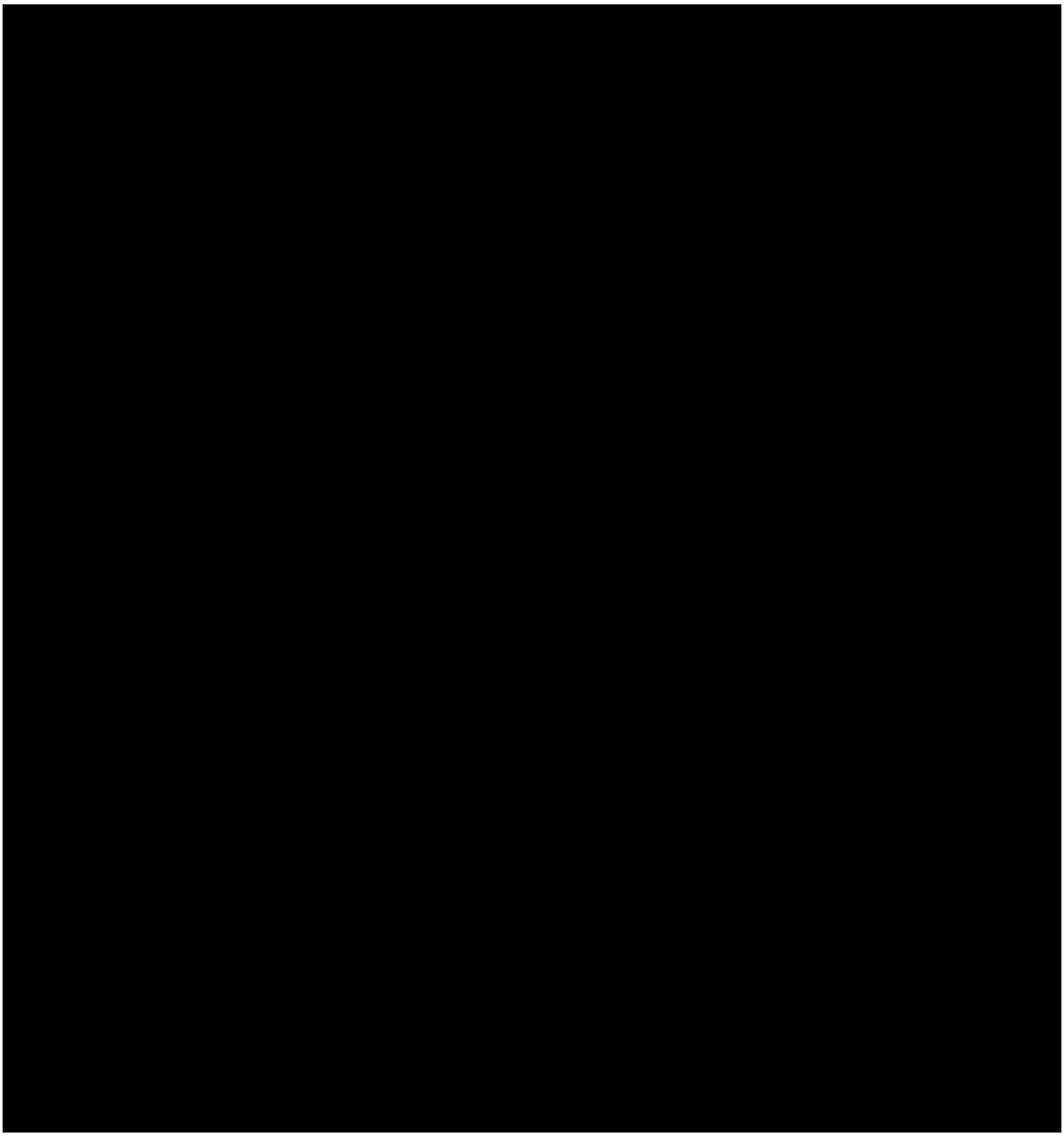
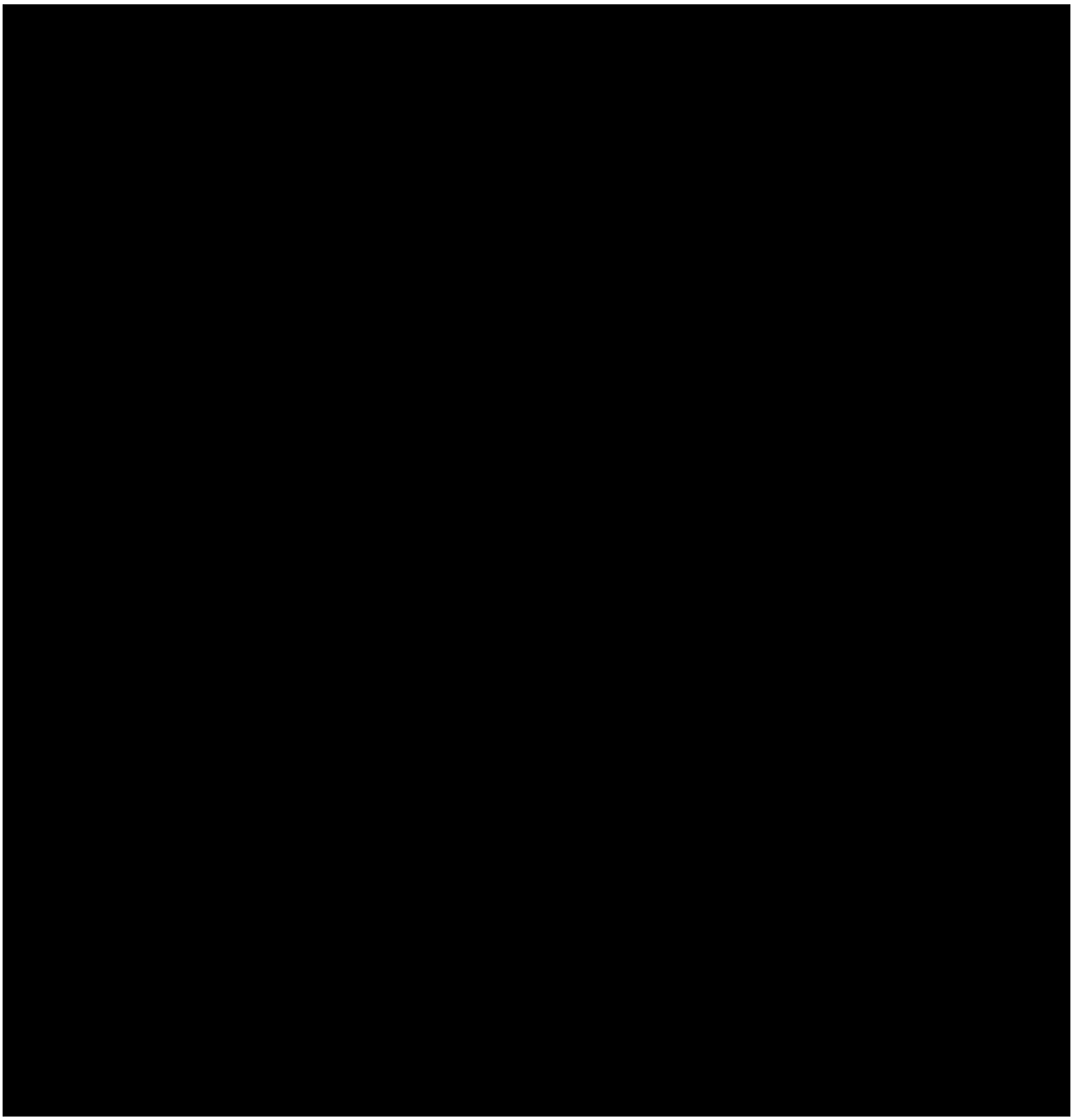


Table 2.7.3.4.2
Equating Summary: Writ 6–8 B/C S501 Online



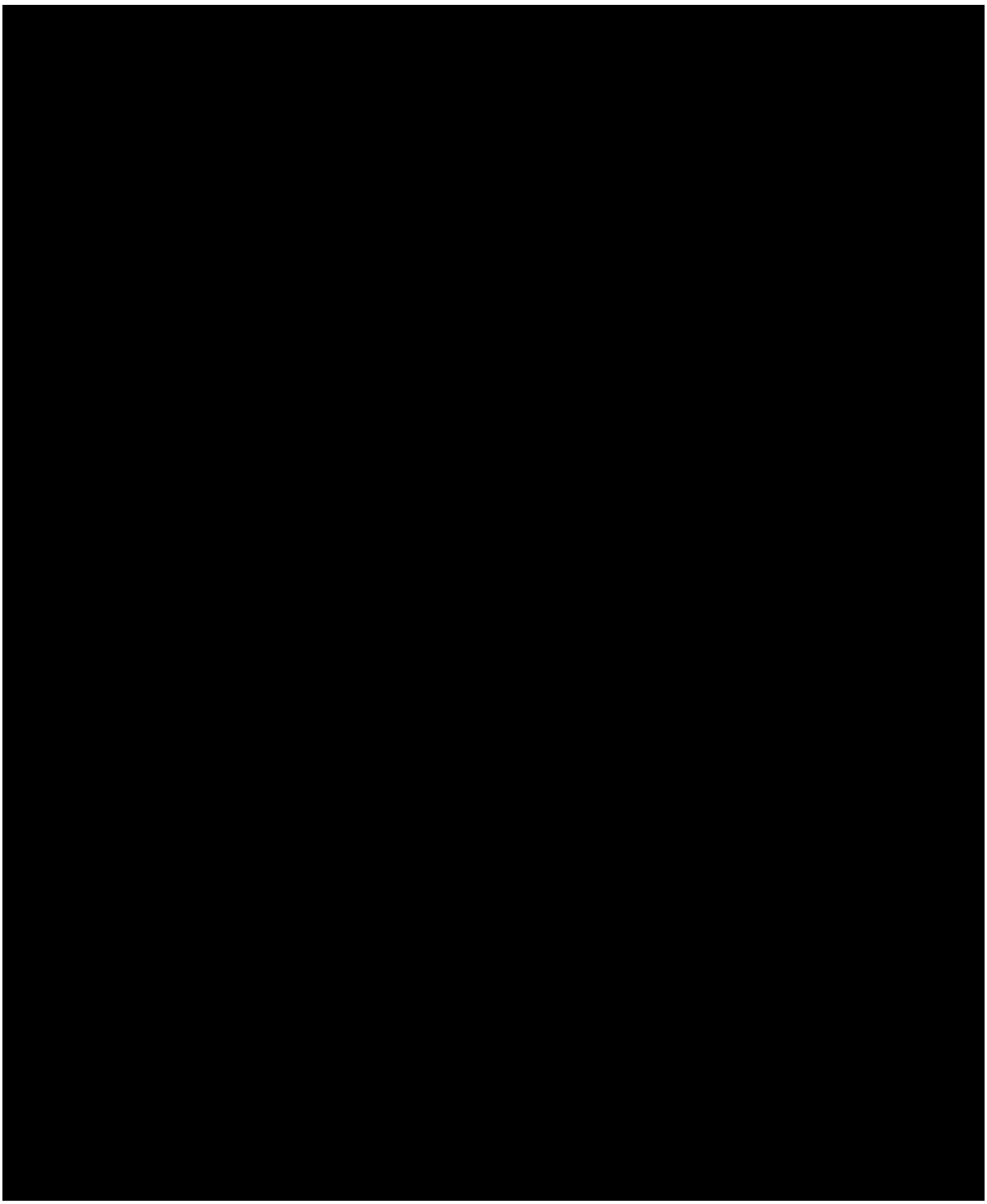
2.7.3.5 *Grades 9–12*



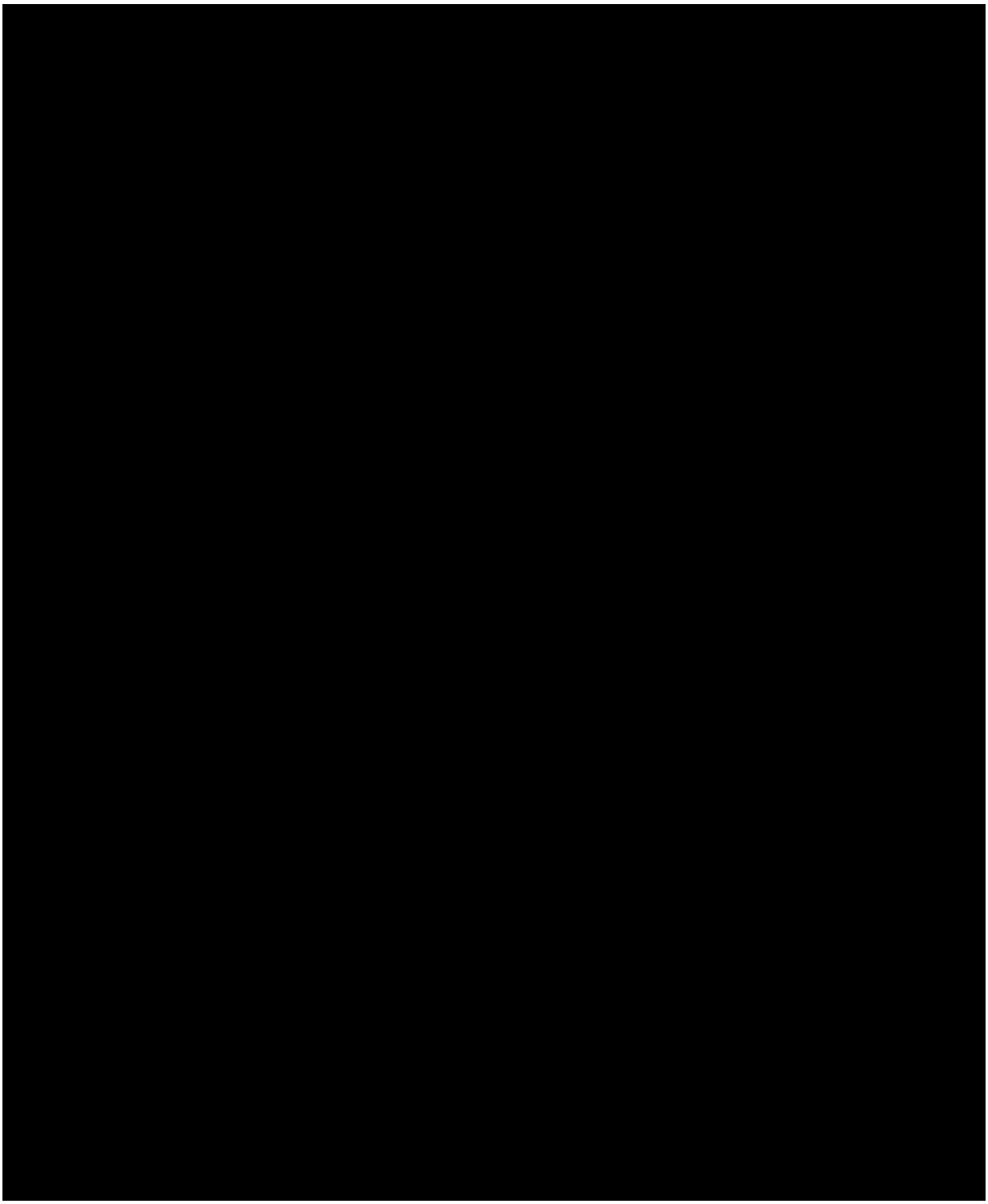


2.7.4 **Speaking**

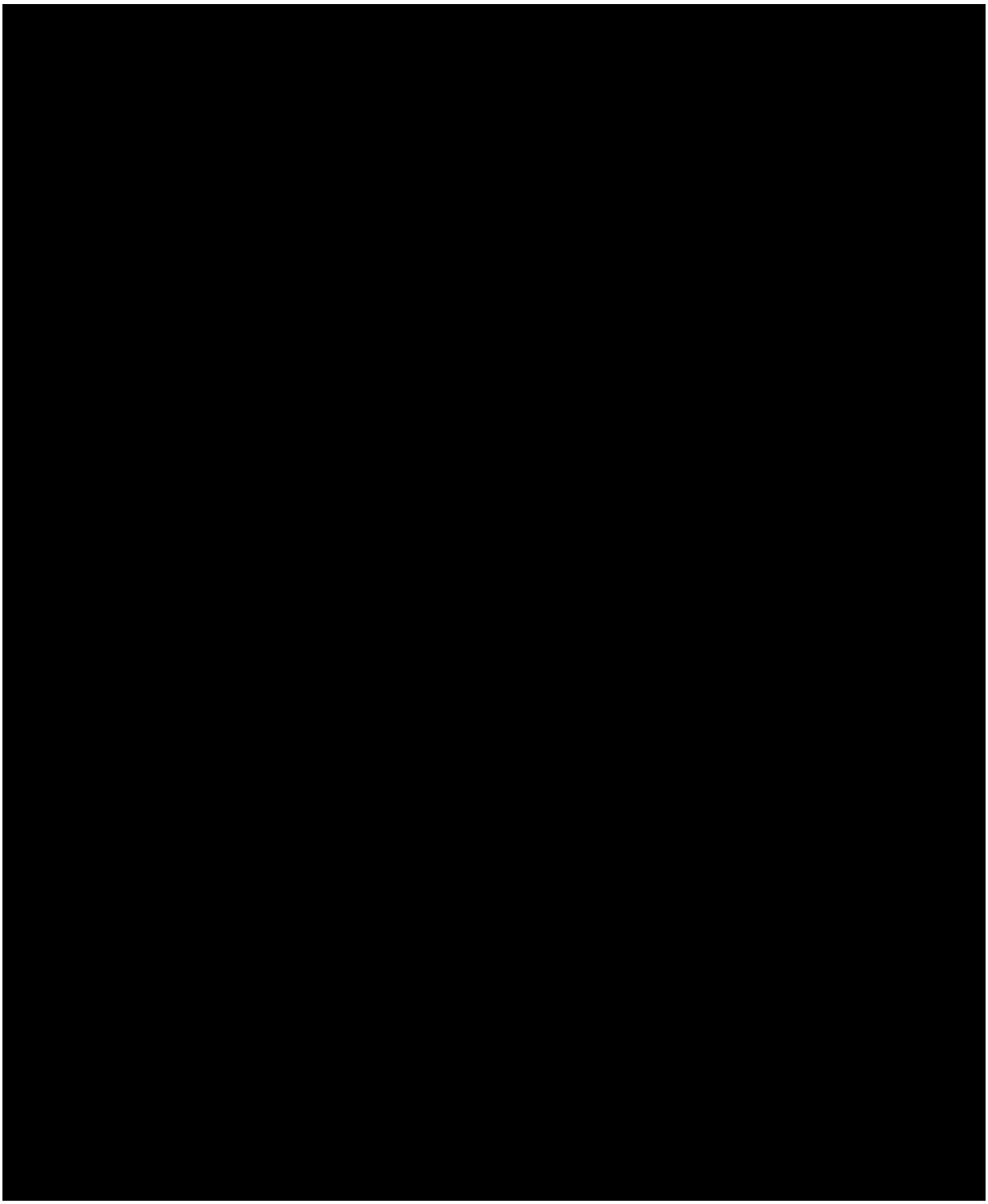
2.7.4.1 *Grade 1*



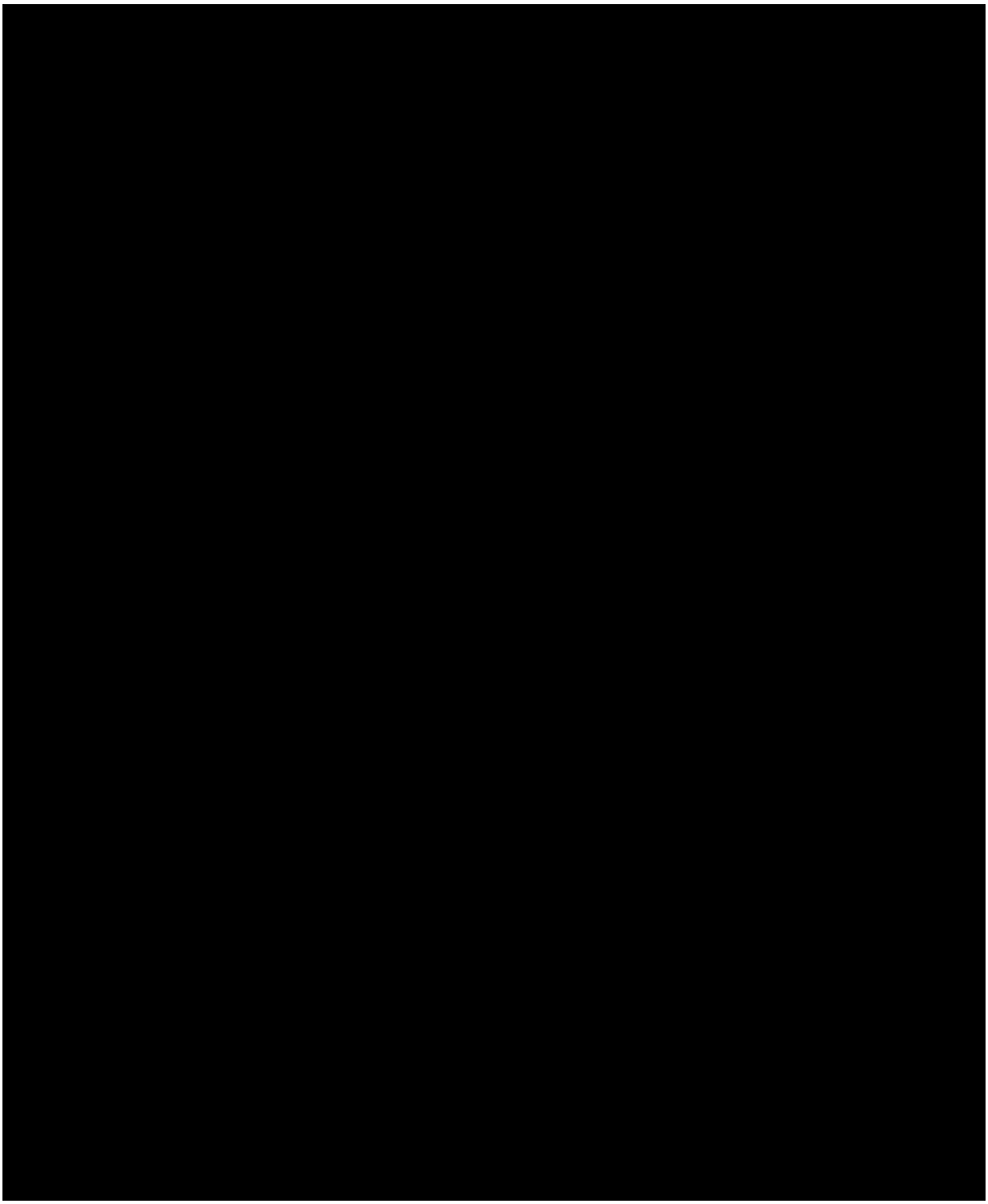
2.7.4.2 *Grades 2–3*



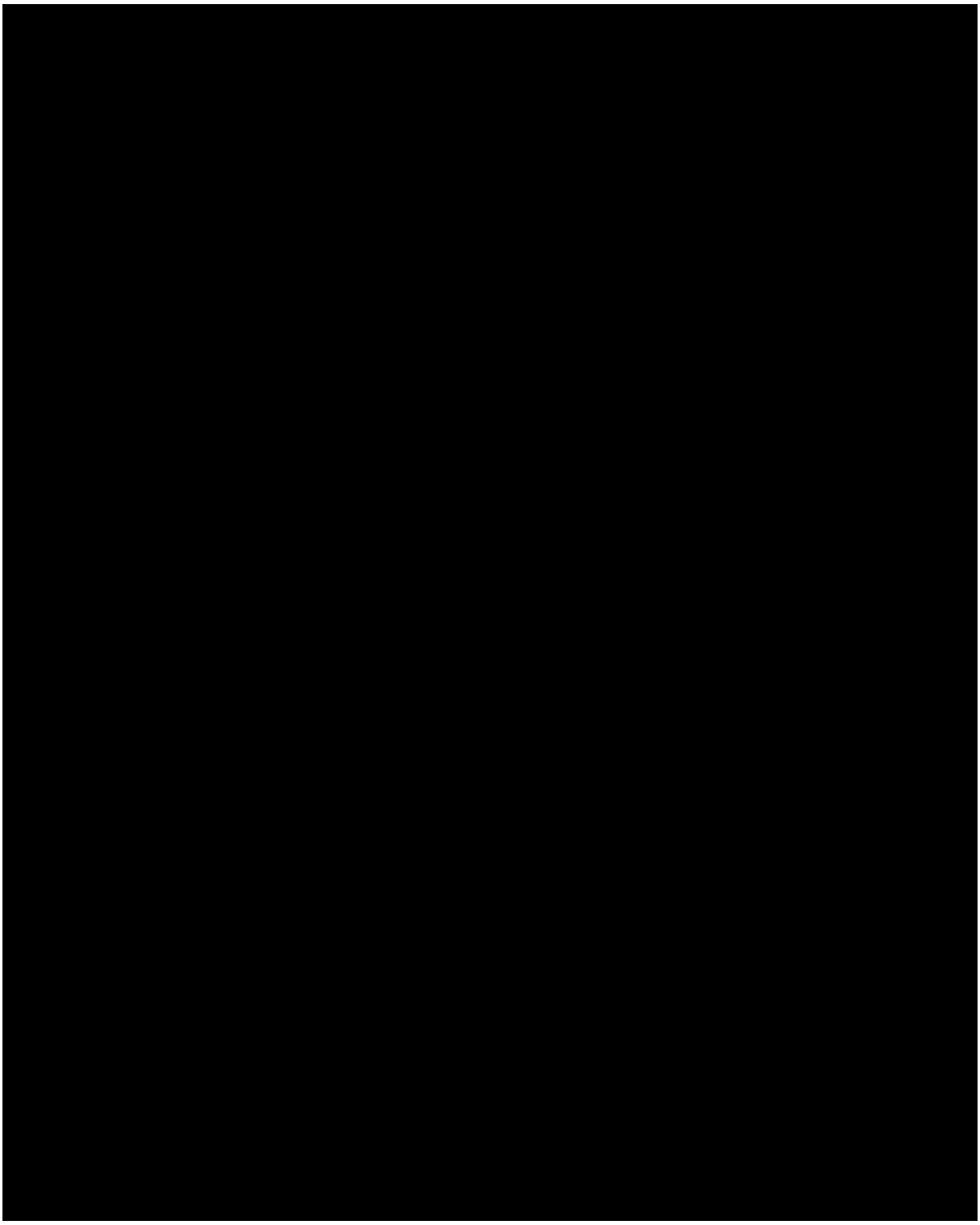
2.7.4.3 *Grades 4–5*



2.7.4.4 *Grades 6–8*



2.7.4.5 *Grades 9–12*



2.8 Test Characteristic Curve

Test characteristic curves (TCC) graphically show the relationship between the ability measure (in logits) on the horizontal axis and the expected raw score or the estimated true score on the vertical axis. For a given ability measure, the corresponding expected raw score can be found via the test characteristic curve. For reporting purposes, ability measures are used to determine students' proficiency levels. Since TCC transforms ability measures to expected raw scores, this representation allows test users to relate student performance to the number of items on the test.

Mathematically, TCC is the sum of all item characteristic functions on the test form (Lord, 1980). Thus, the TCC depends on the item characteristic functions (Lord, 1980). The shape of TCC depends on several factors, including the number and the characteristics of items, the item response theory model used, and the values of the item parameters. Because of this, there is no explicit formula for TCC, and there are no parameters for the curve.

Listening and Reading Online ACCESS tests are presented in a multistage adaptive format and are not fixed test forms. Therefore, t is not appropriate to present TCC for these domains.

For the Writing and Speaking tests, which consist of polytomous tasks, the shape of the TCC is also affected by the values of the item category parameters. Since polytomous tasks have more score categories than multiple-choice items, each task has a wide range of values on the proficiency scale. The adjacent category boundaries are sometimes far apart as a result. In this situation, the TCC will have a less smooth curve or a small plateau in the area between the adjacent category boundaries. This pattern can be observed in Writing and Speaking, where the TCC may not form a perfect "S" shape. Such a pattern is also observed in other tests with polytomous items, such as the National Assessment of Educational Progress Writing assessment (Muraki, 1993). Conversely, the closer the adjacent category boundaries are, the smoother the rise of the TCC will be along the ability levels.

There are five vertical lines in each of the TCC plots indicating the five cut scores 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 (PLs 1–6) for the domain being tested. 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, however, gives an indication of how many points must be earned to be placed into a WIDA proficiency level.

In addition to the TCC by tier, TCCs across tier for the grade level-cluster are plotted on the same graph. Since each tier has different numbers of expected raw score points, it is not appropriate to compare the expected raw score points for the same proficiency measure between tiers. It is, however, informative to compare where the slopes are the steepest, which corresponds to the ability range where the best measurement information is provided. For example, the across-tier TCC for Writing Grade 1 showed that the Writing Tier A form provides better measurement at around ability measures of -2.5 and 3.0, whereas the Writing Grade 1 Tier B/C form provides better measurement at around measures of -2.0 and 3.5. In addition, it is

informative to compare the area under the curve for the TCC of each tier form. For example, the Grade 1 Tier A curve covers an area of lower ability range than the Grade 1 Tier B/C curve, especially at the very low end of the ability range. Consistent with the purposes of the test design, there is also considerable overlap between the areas covered by the two forms.

2.8.1 Listening

The ACCESS 2.0 Online Listening test is a multistage adaptive assessment. As students do not all take the same set of items in the test, no test characteristic curve is presented.

2.8.2 Reading

The ACCESS 2.0 Online Reading test is a multistage adaptive assessment. As students do not all take the same set of items in the test, no test characteristic curve is presented.

2.8.3 Writing

2.8.3.1 *Grade 1*

Figure 2.8.3.1.1
Test Characteristic Curve: Writ 1 A S501 Online

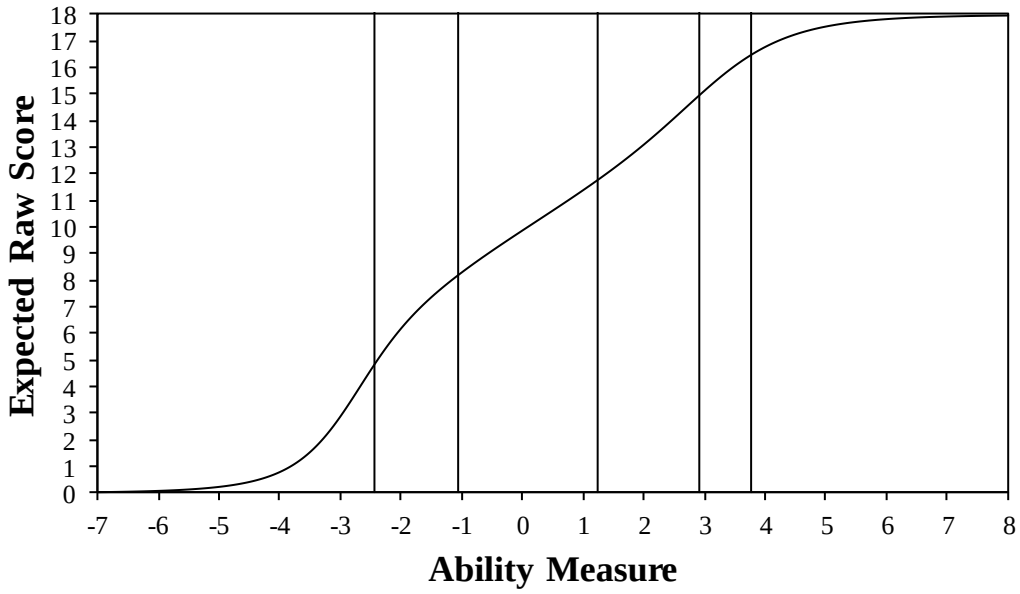


Figure 2.8.3.1.2
Test Characteristic Curve: Writ 1 B/C S501 Online

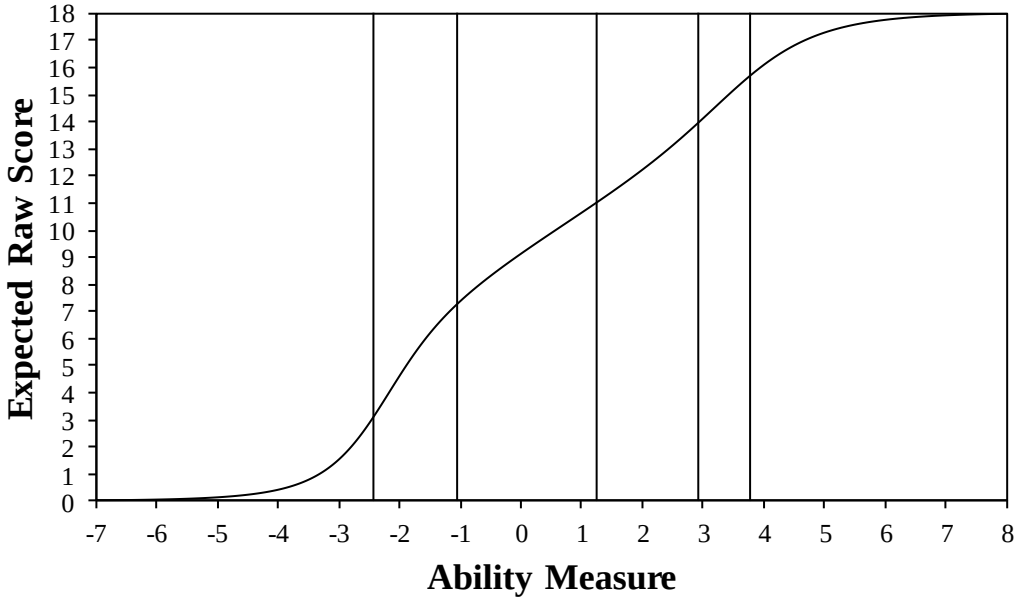
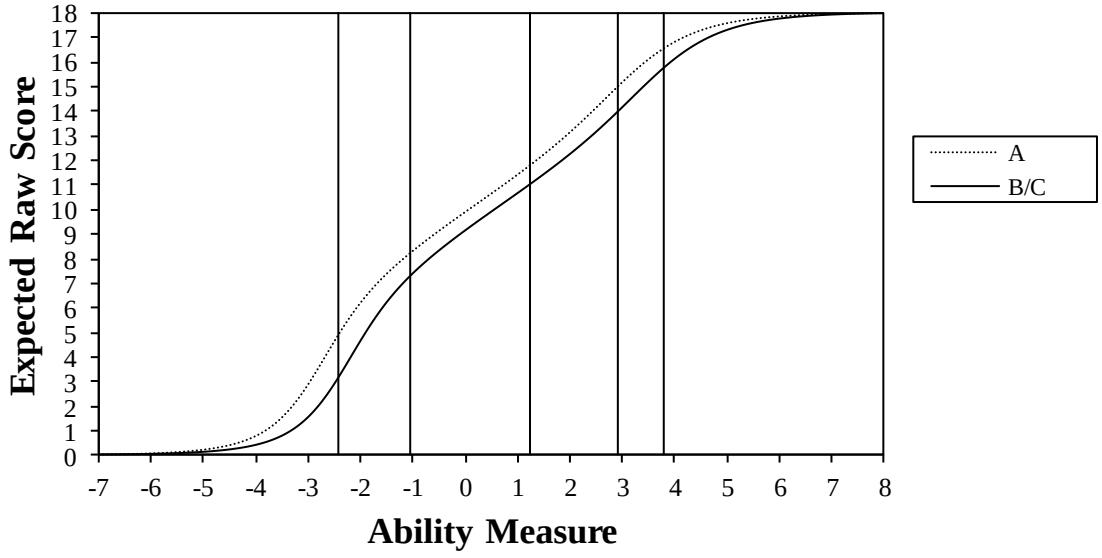


Figure 2.8.3.1.3
Test Characteristic Curve: Writ 1 S501 Online



2.8.3.2 *Grades 2–3*

Figure 2.8.3.2.1
Test Characteristic Curve: Writ 2-3 A S501 Online

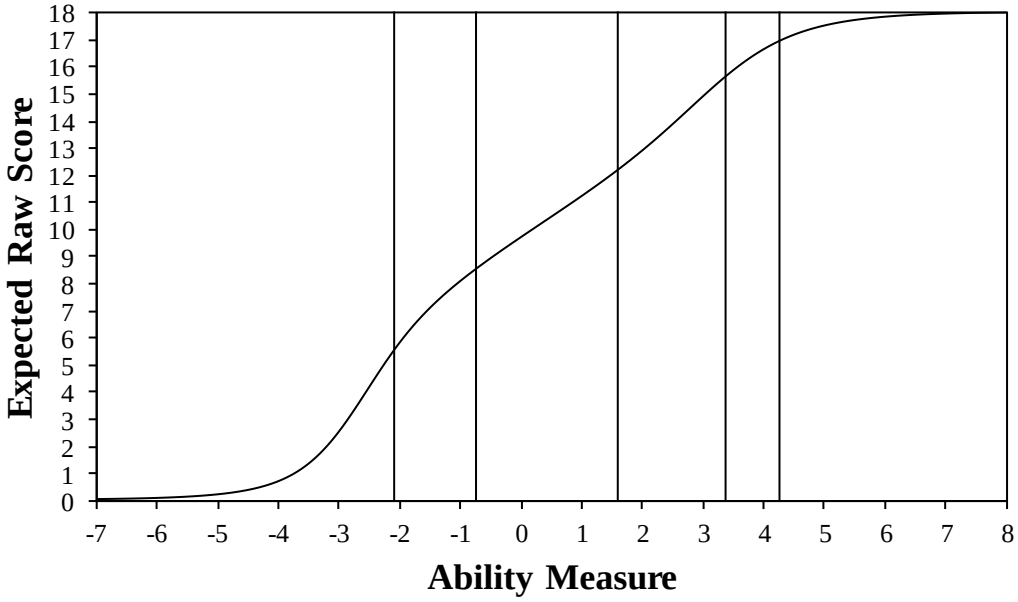


Figure 2.8.3.2.2
Test Characteristic Curve: Writ 2-3 B/C S501 Online

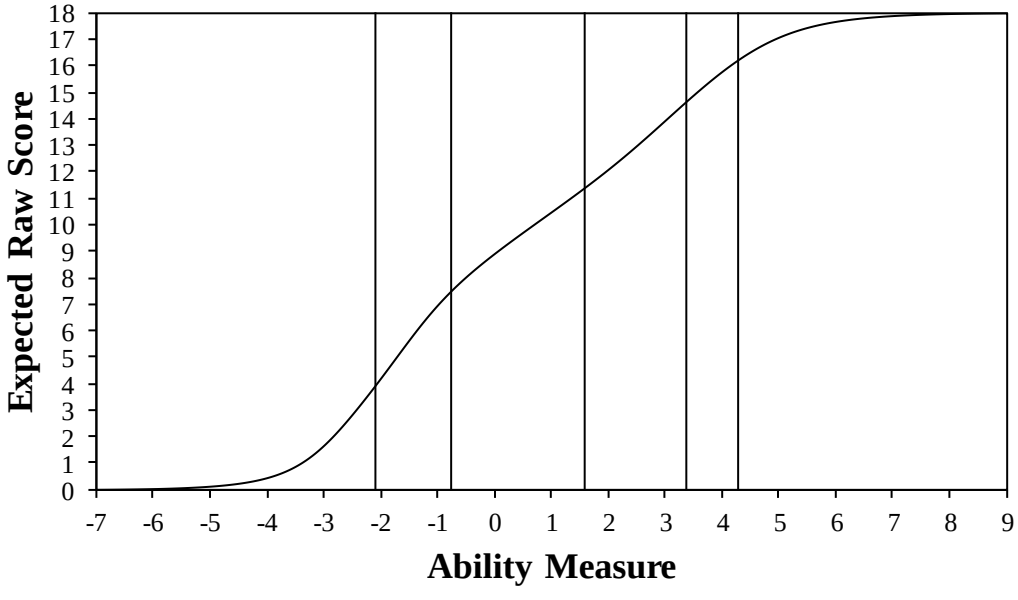
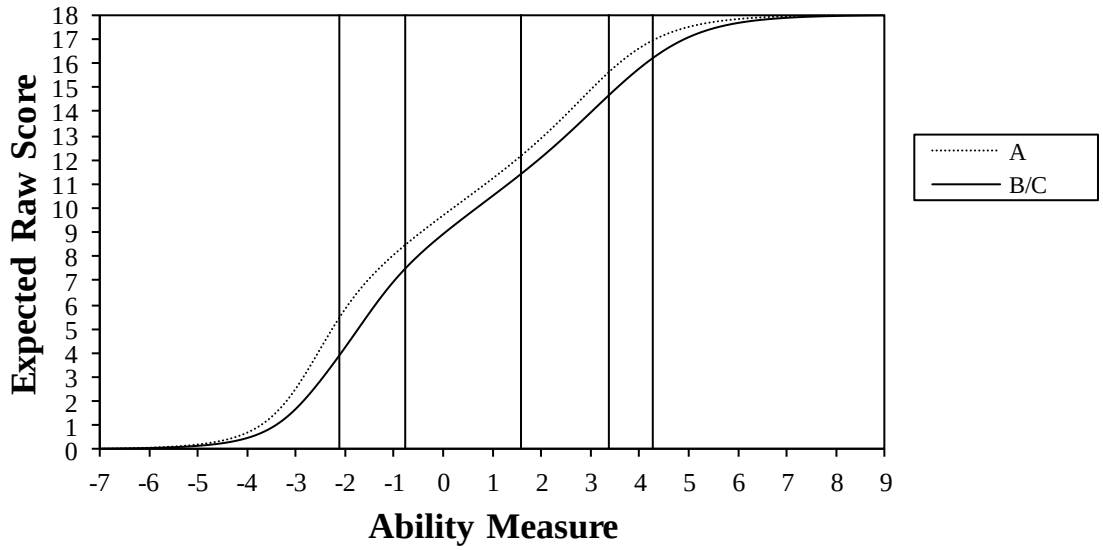


Figure 2.8.3.2.3
Test Characteristic Curve: Writ 2-3 S501 Online



2.8.3.3 *Grades 4–5*

Figure 2.8.3.3.1
Test Characteristic Curve: Writ 4-5 A S501 Online

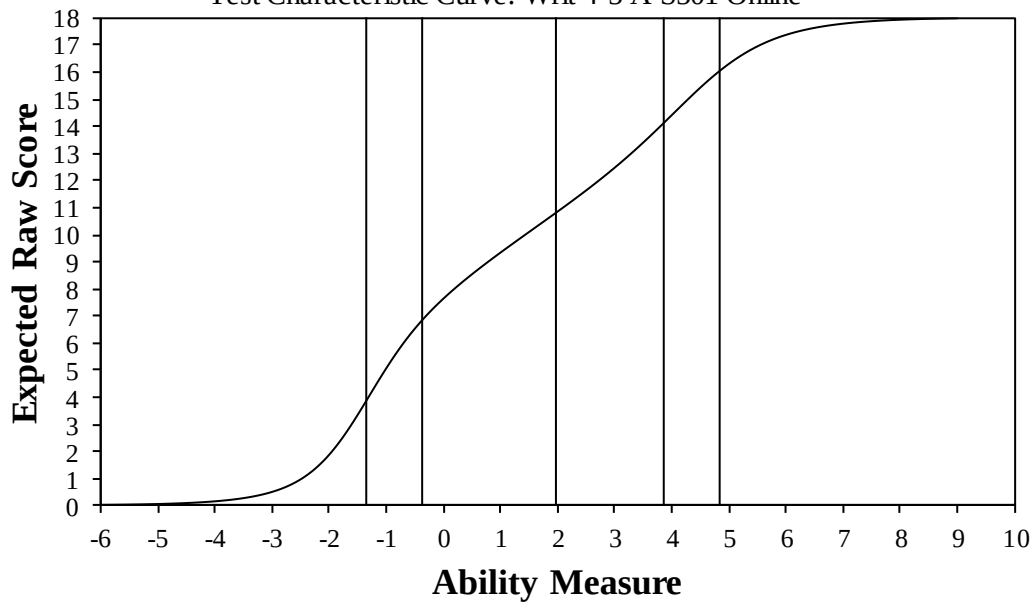


Figure 2.8.3.3.2
Test Characteristic Curve: Writ 4-5 B/C S501 Online

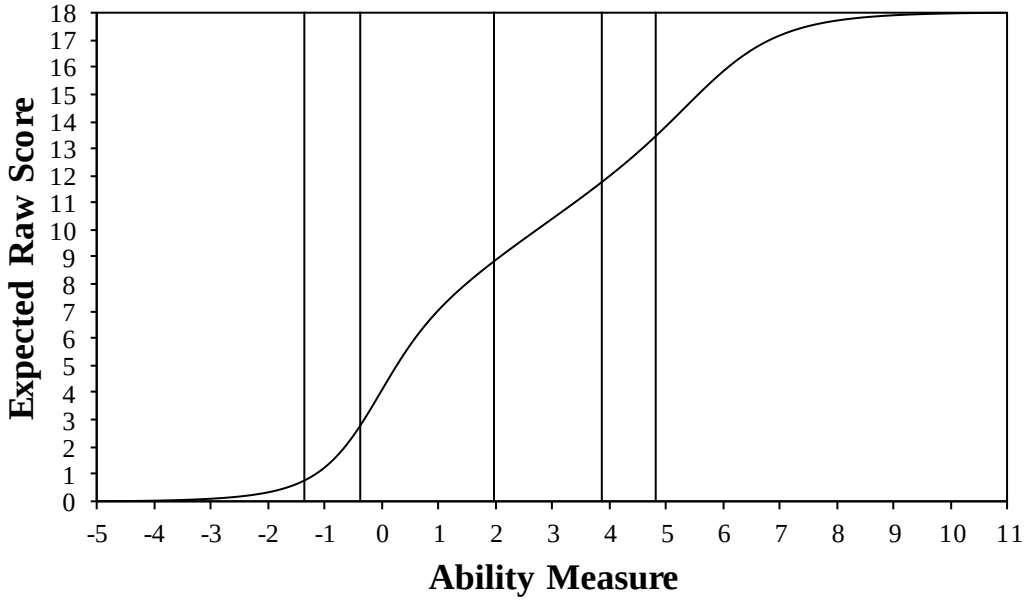
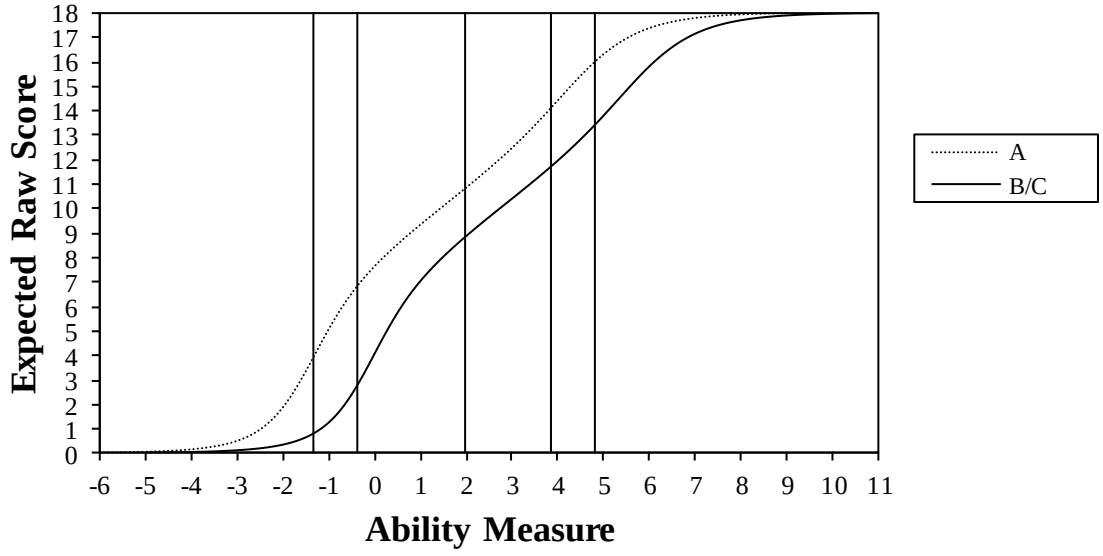


Figure 2.8.3.3.3
Test Characteristic Curve: Writ 4-5 S501 Online



2.8.3.4 *Grades 6–8*

Figure 2.8.3.4.1
Test Characteristic Curve: Writ 6-8 A S501 Online

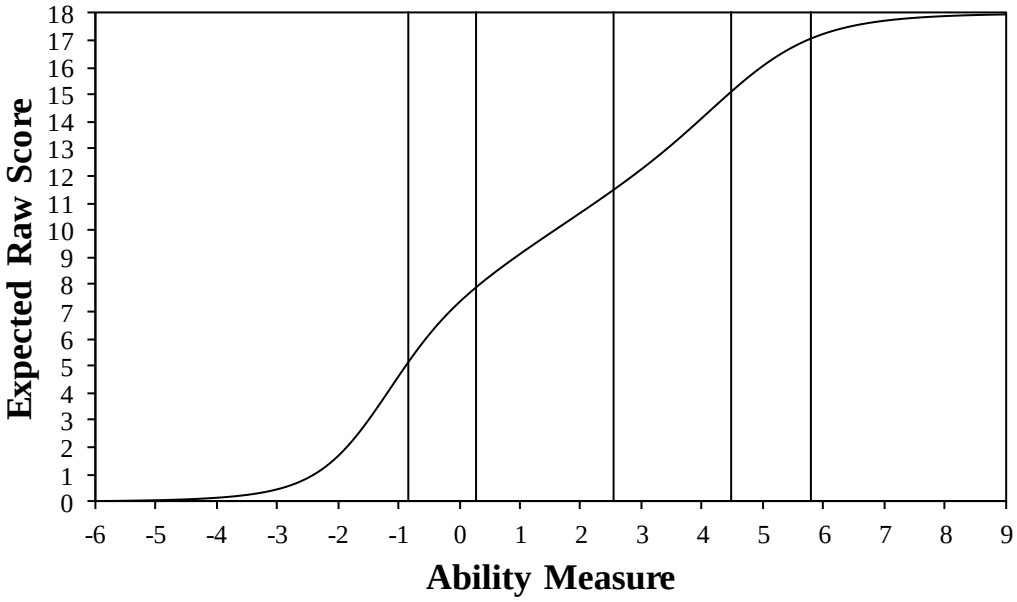


Figure 2.8.3.4.2
Test Characteristic Curve: Writ 6-8 B/C S501 Online

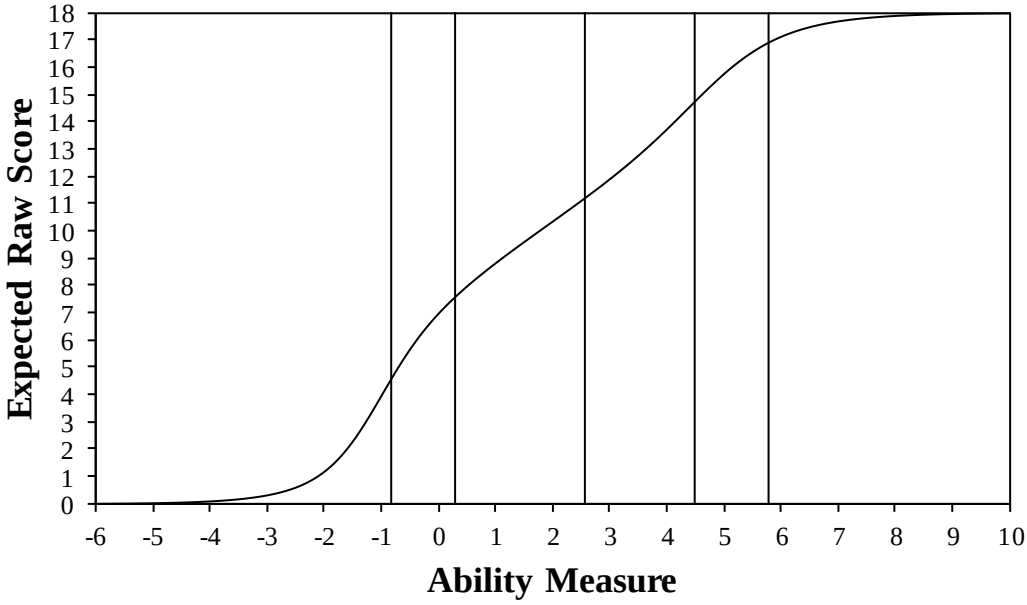
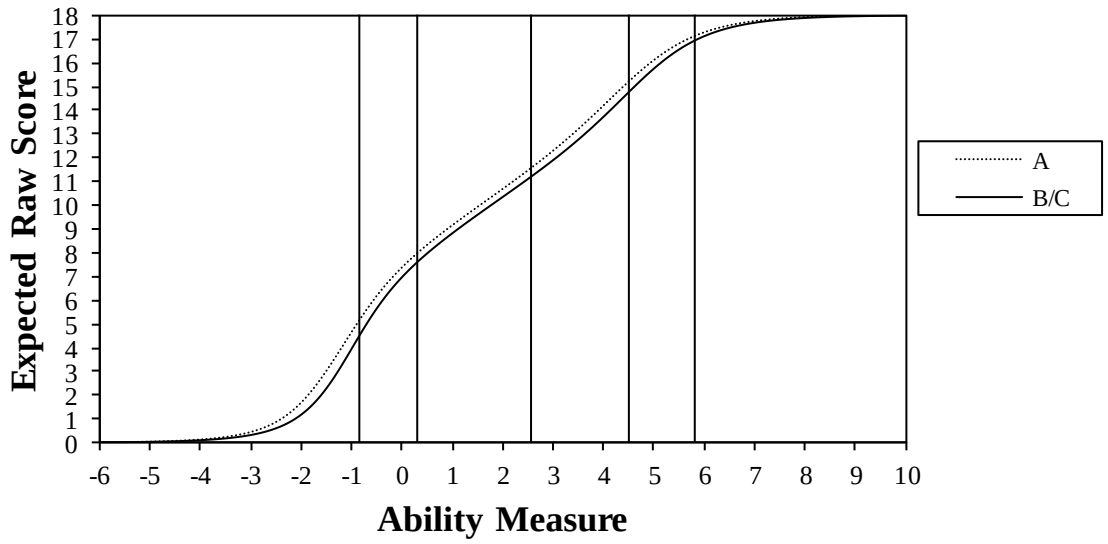


Figure 2.8.3.4.3
Test Characteristic Curve: Writ 6-8 S501 Online



2.8.3.5 *Grades 9–12*

Figure 2.8.3.5.1
Test Characteristic Curve: Writ 9-12 A S501 Online

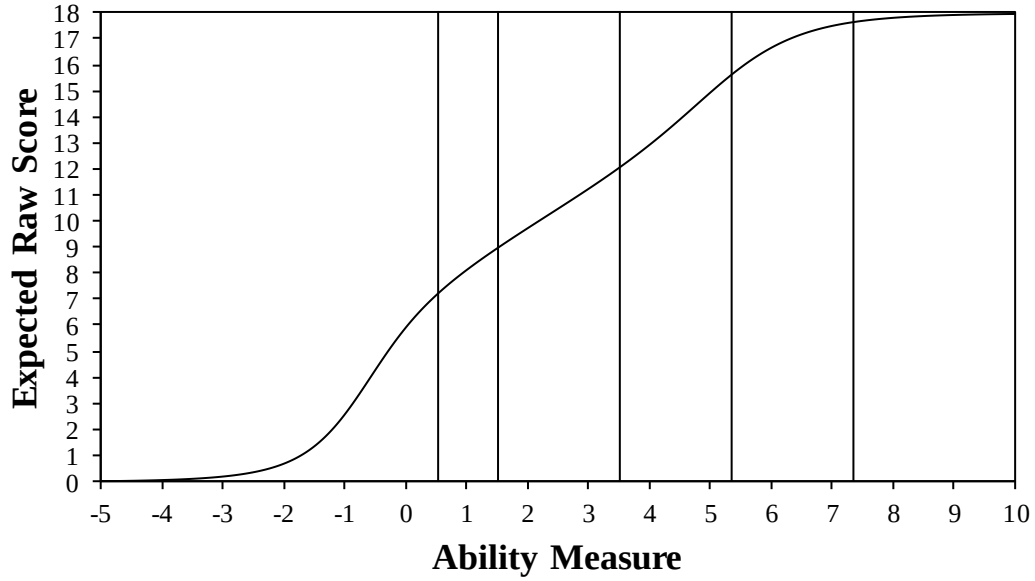


Figure 2.8.3.5.2
Test Characteristic Curve: Writ 9-12 B/C S501 Online

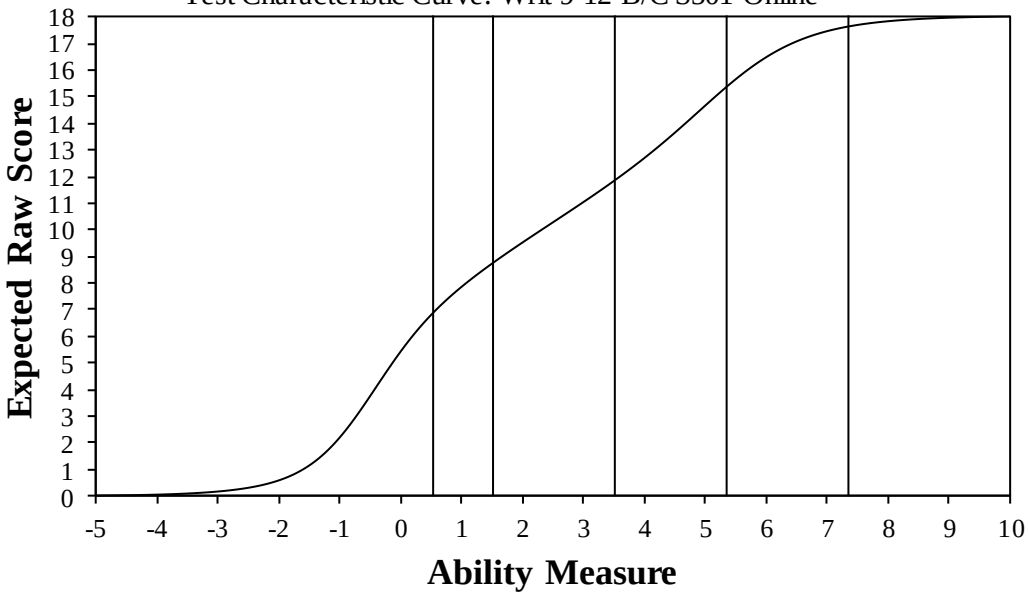
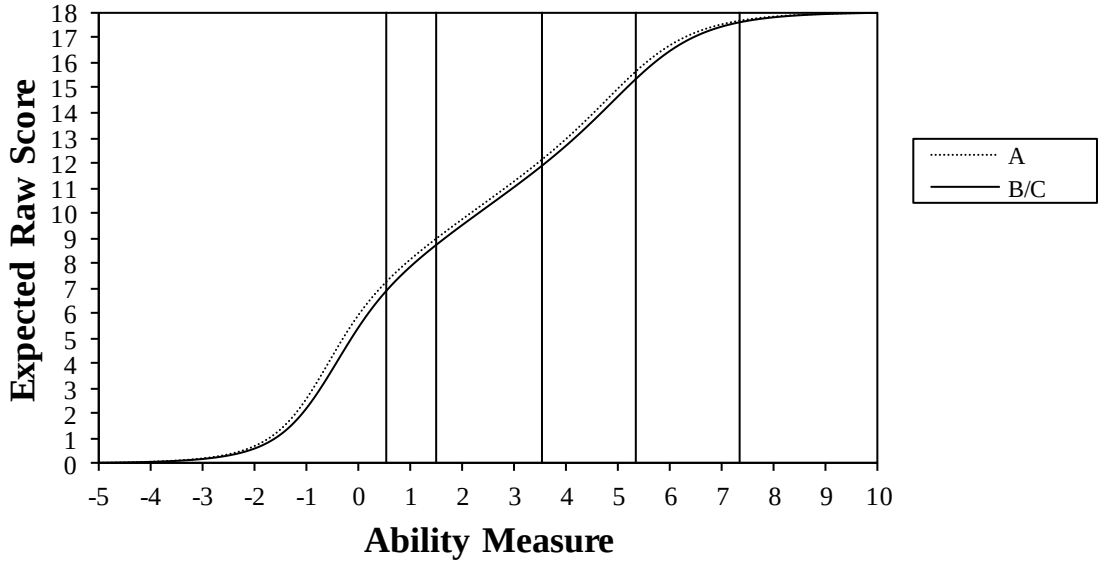


Figure 2.8.3.5.3
Test Characteristic Curve: Writ 9-12 S501 Online



2.8.4 **Speaking**

2.8.4.1 *Grade 1*

Figure 2.8.4.1.1
Test Characteristic Curve: Spek 1 Pre-A S501 Online

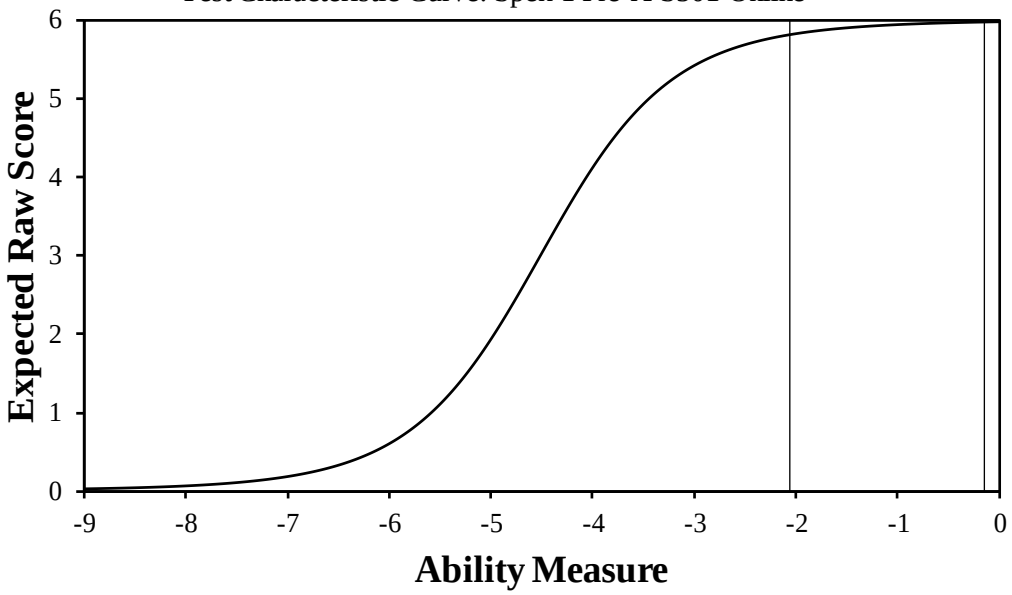


Figure 2.8.4.1.2
Test Characteristic Curve: Spek 1 A S501 Online

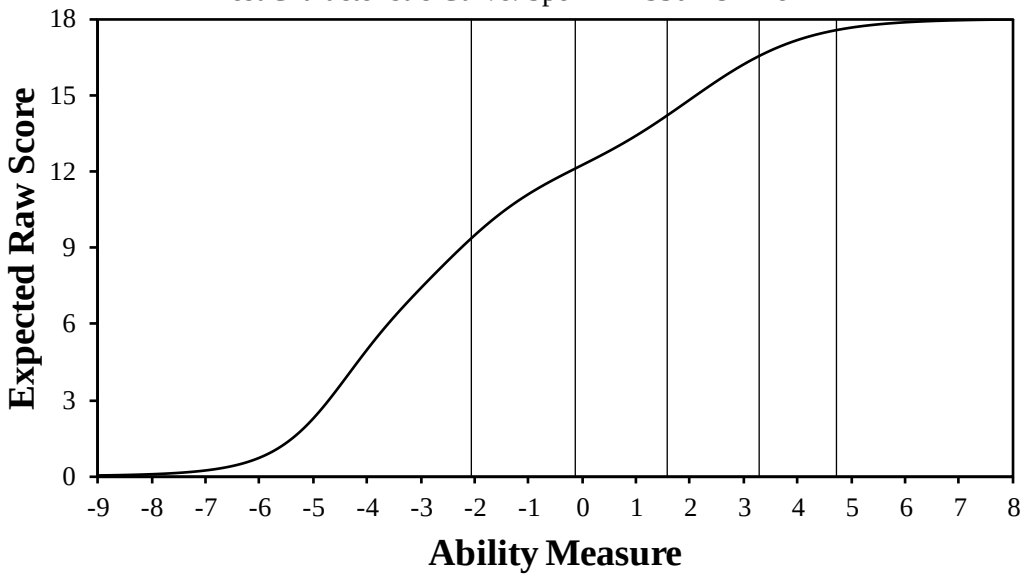


Figure 2.8.4.1.3
Test Characteristic Curve: Spek 1 B/C S501 Online

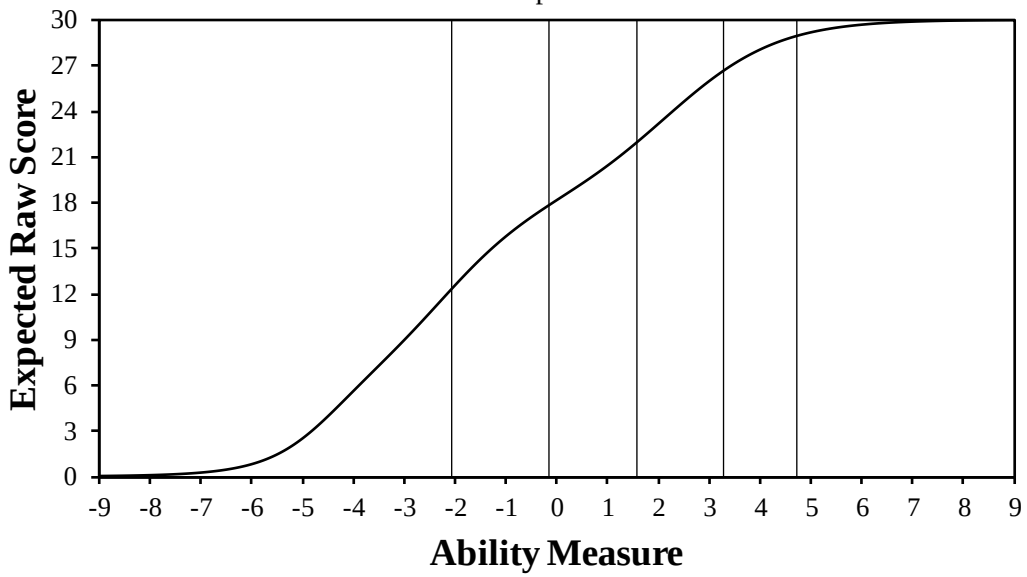
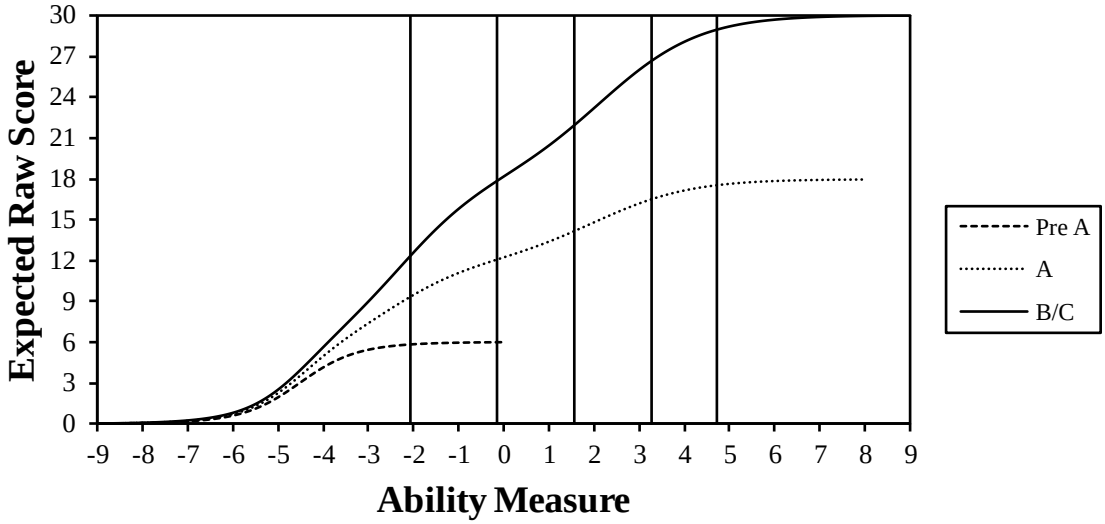


Figure 2.8.4.1.4
Test Characteristic Curve: Spek 1 S501 Online



2.8.4.2 *Grades 2–3*

Figure 2.8.4.2.1
Test Characteristic Curve: Spek 2-3 Pre-A S501 Online

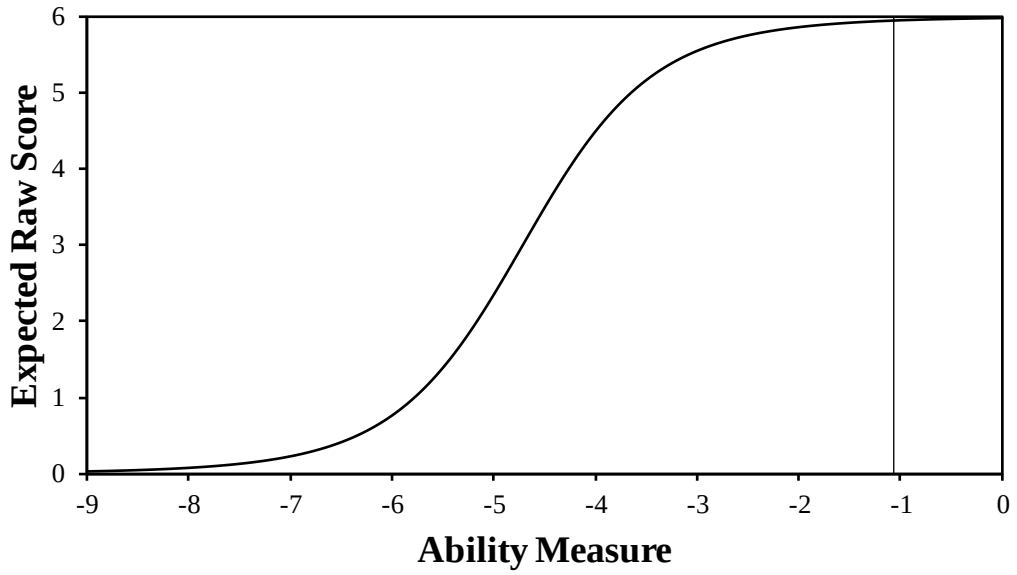


Figure 2.8.4.2.2
Test Characteristic Curve: Spek 2-3 A S501 Online

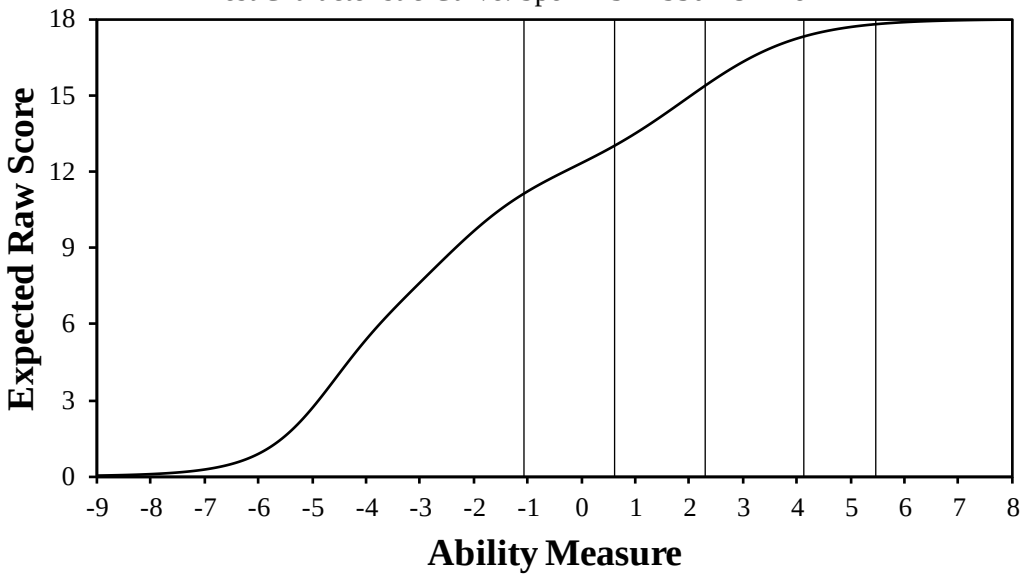


Figure 2.8.4.2.3
Test Characteristic Curve: Spek 2-3 B/C S501 Online

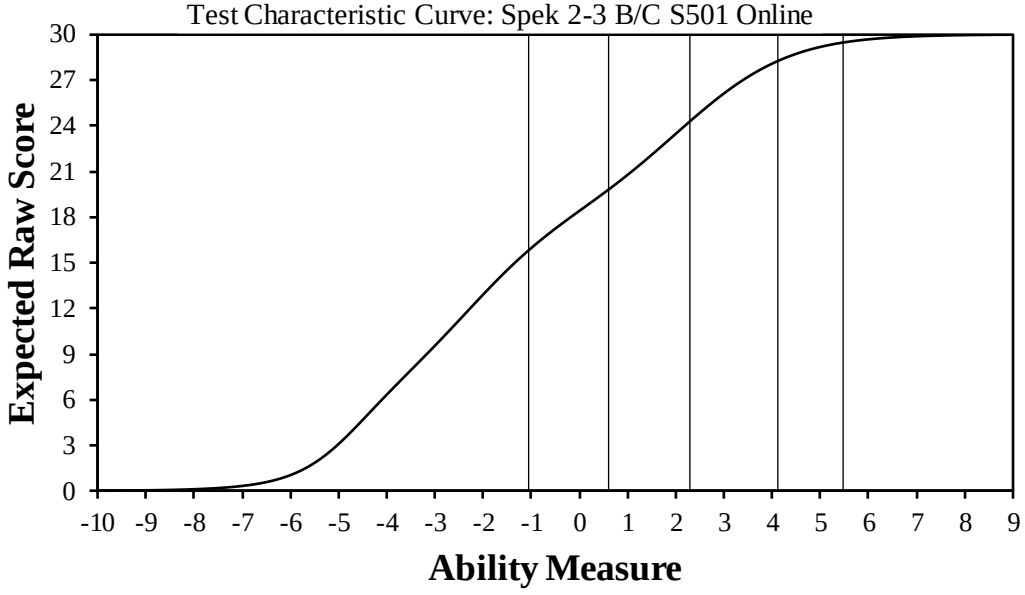
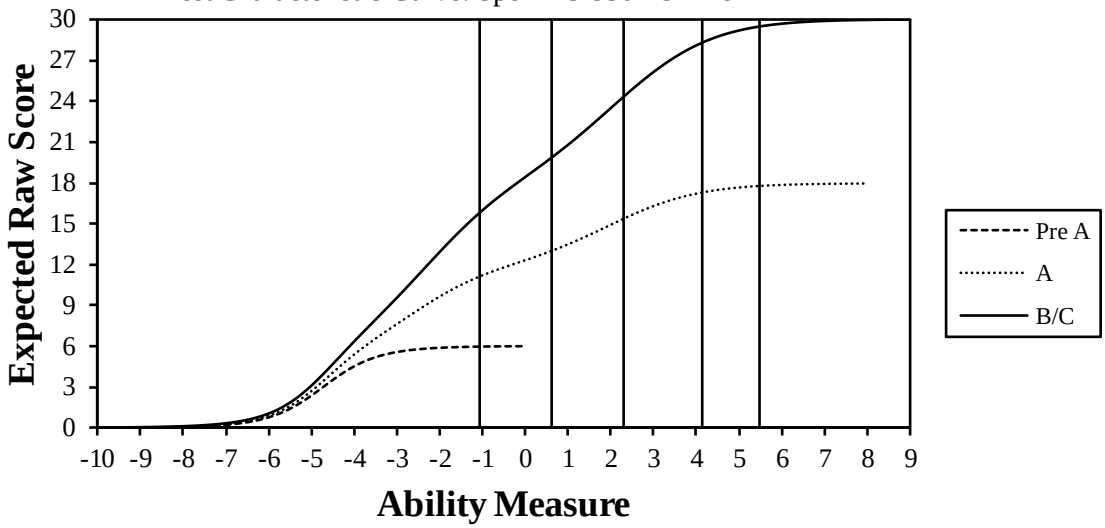


Figure 2.8.4.2.4
Test Characteristic Curve: Spek 2-3 S501 Online



2.8.4.3 *Grades 4–5*

Figure 2.8.4.3.1
Test Characteristic Curve: Spek 4-5 Pre-A S501 Online

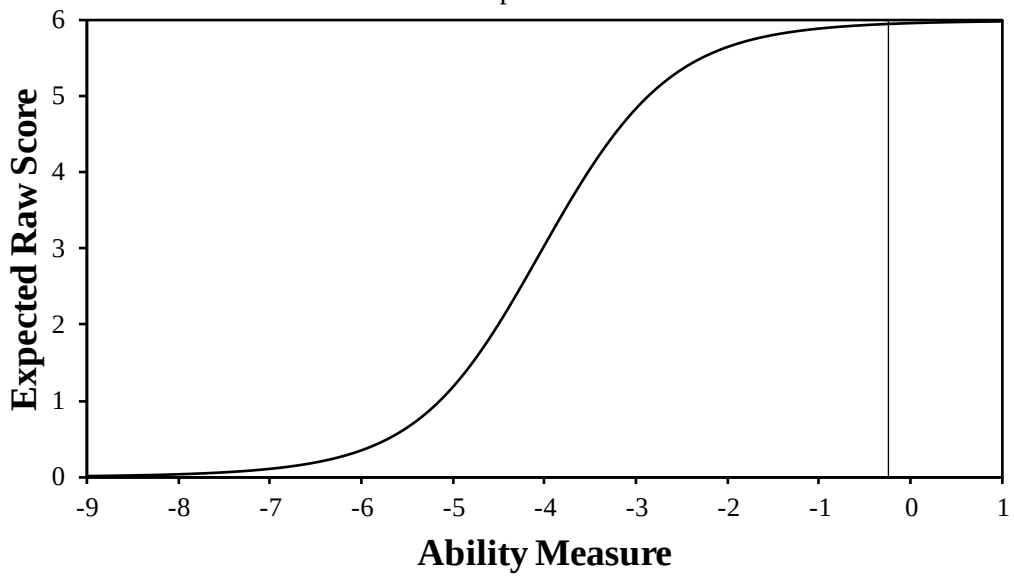


Figure 2.8.4.3.2
Test Characteristic Curve: Spek 4-5 A S501 Online

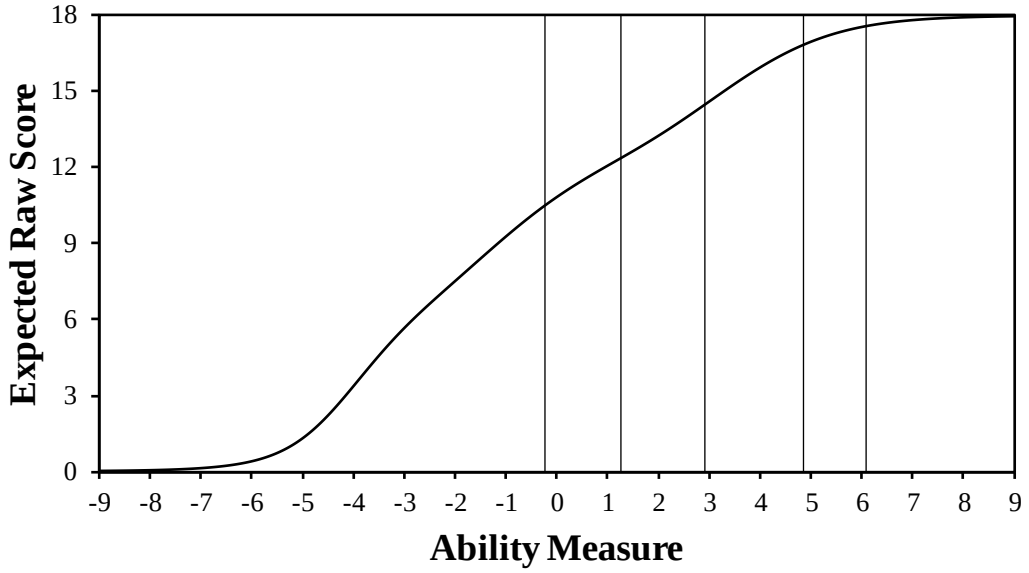


Figure 2.8.4.3.3
Test Characteristic Curve: Spek 4-5 B/C S501 Online

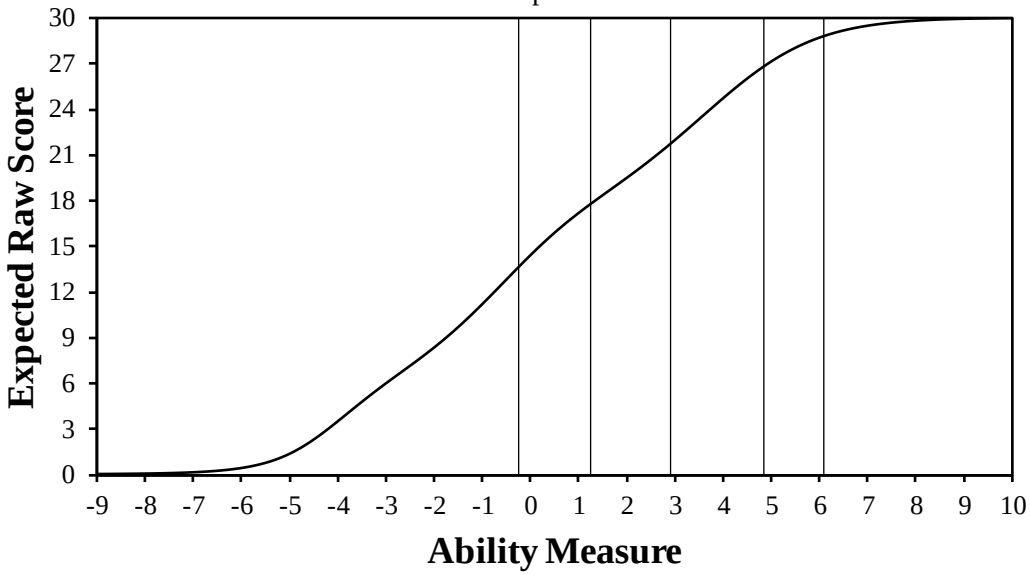
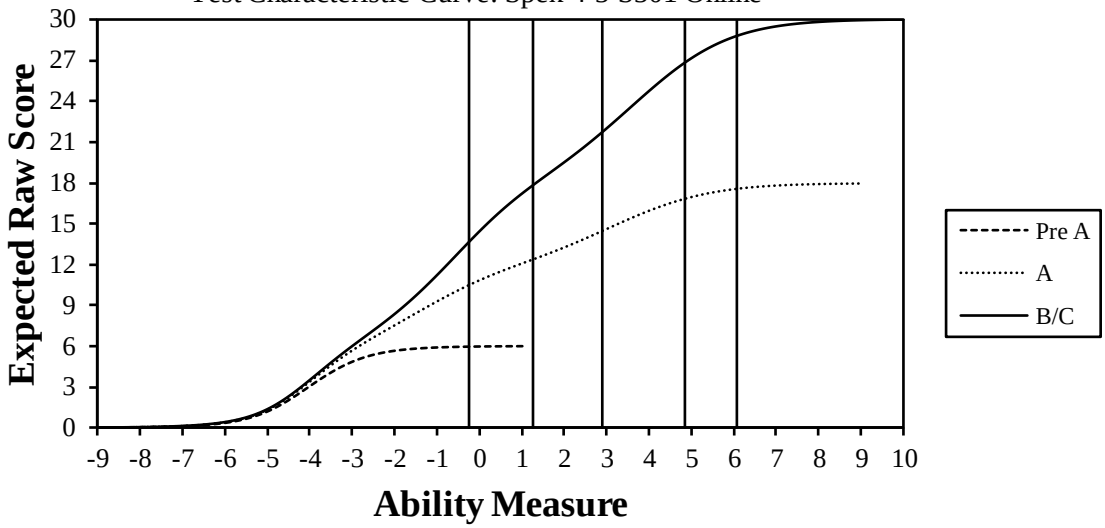


Figure 2.8.4.3.4
Test Characteristic Curve: Spek 4-5 S501 Online



2.8.4.4 *Grades 6–8*

Figure 2.8.4.4.1
Test Characteristic Curve: Spek 6-8 Pre-A S501 Online

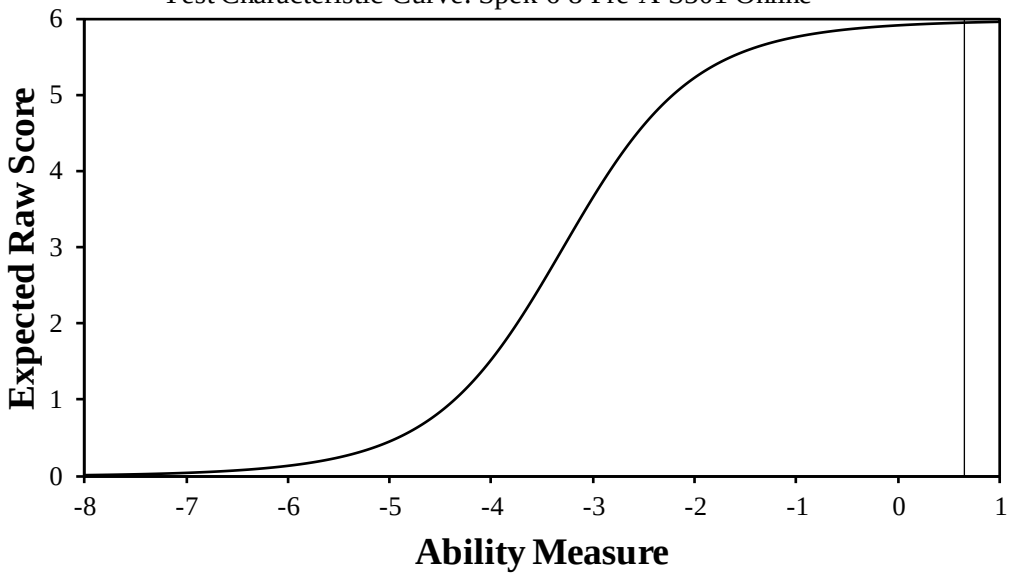


Figure 2.8.4.4.2
Test Characteristic Curve: Spek 6-8 A S501 Online

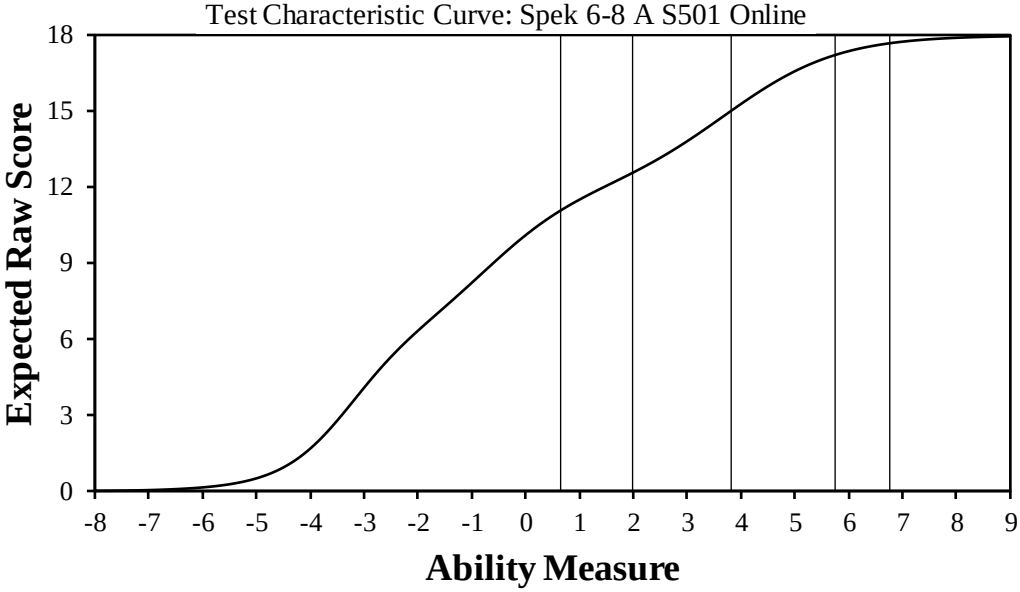


Figure 2.8.4.4.3
Test Characteristic Curve: Spek 6-8 B/C S501 Online

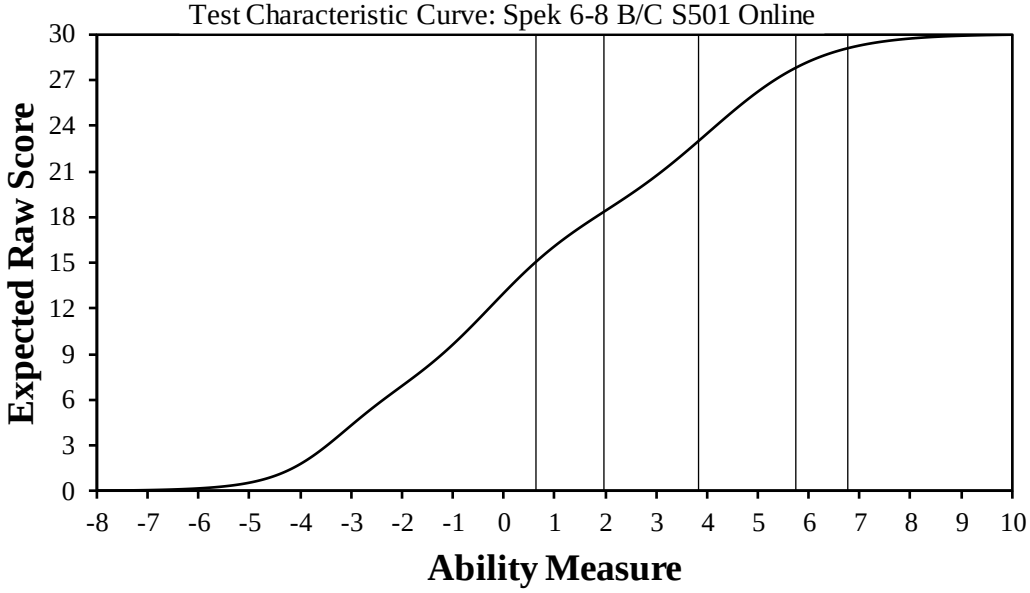
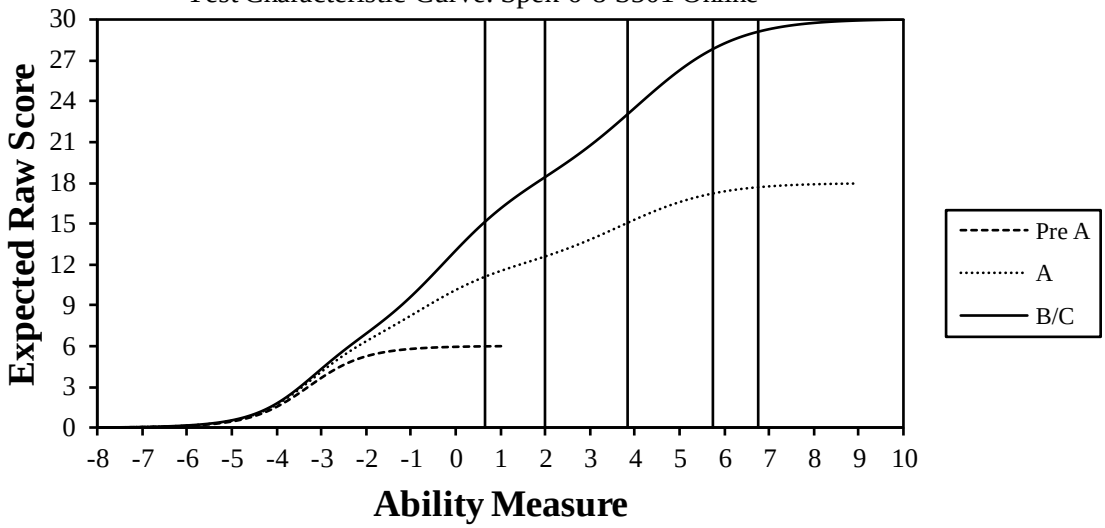


Figure 2.8.4.4.4
Test Characteristic Curve: Spek 6-8 S501 Online



2.8.4.5 *Grades 9–12*

Figure 2.8.4.5.1
Test Characteristic Curve: Spek 9-12 Pre-A S501 Online

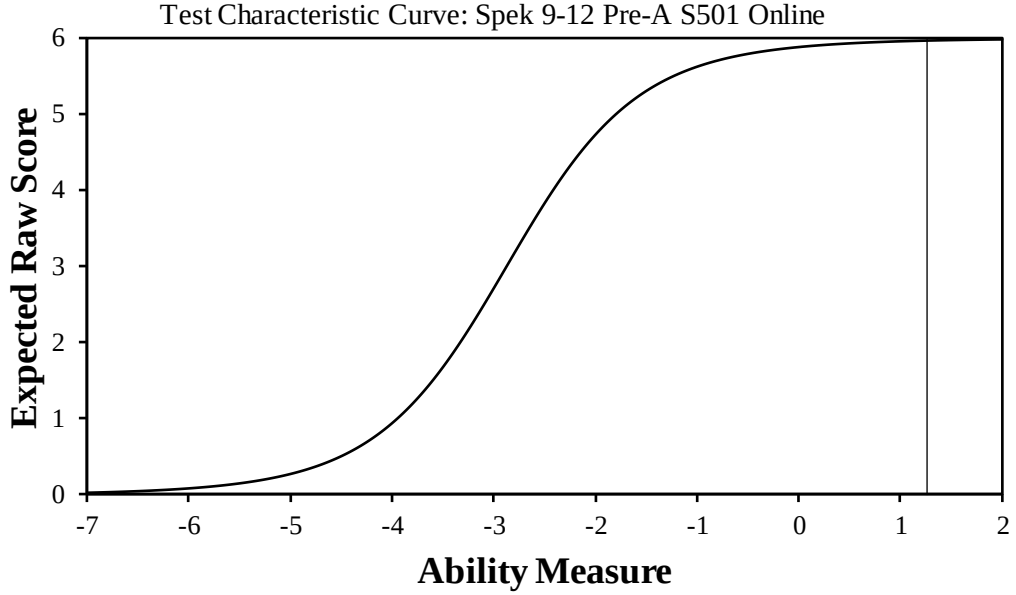


Figure 2.8.4.5.2
Test Characteristic Curve: Spek 9-12 A S501 Online

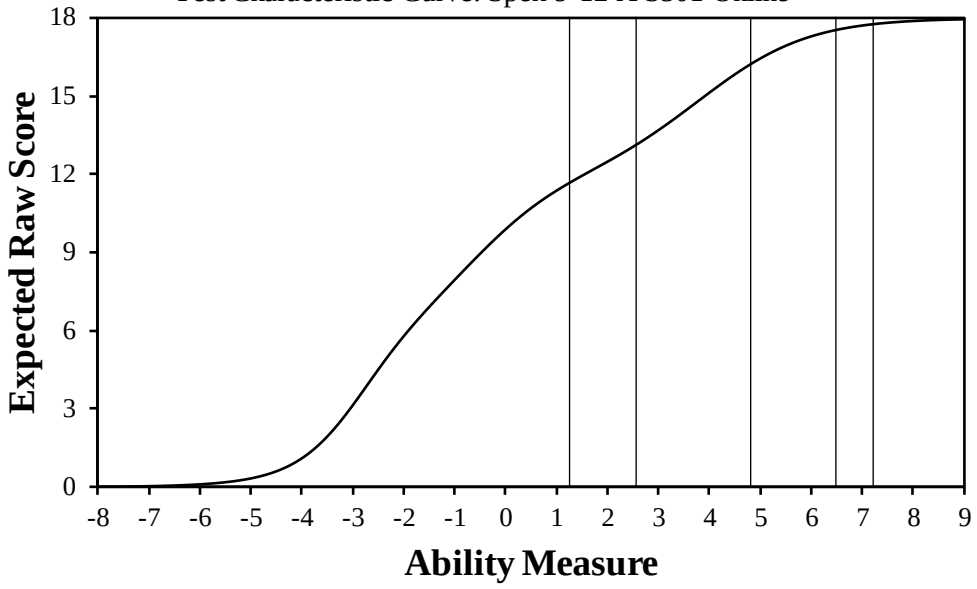


Figure 2.8.4.5.3
Test Characteristic Curve: Spek 9-12 B/C S501 Online

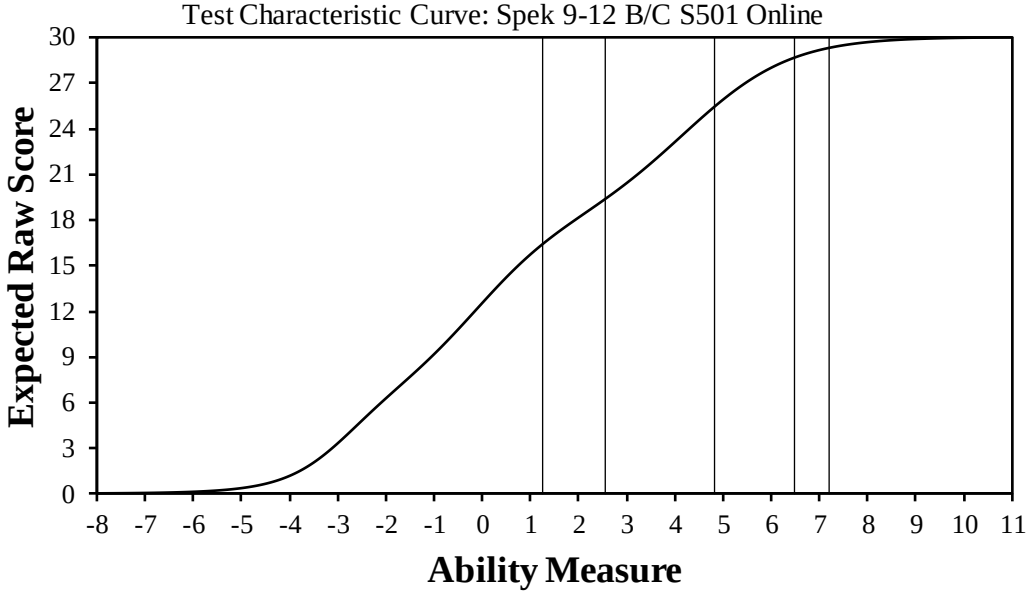
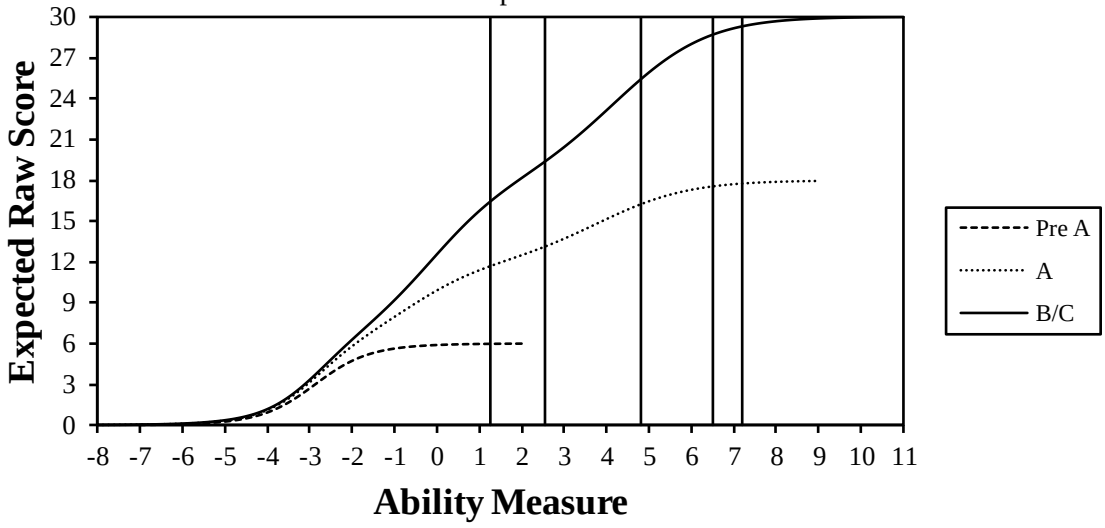


Figure 2.8.4.5.4
Test Characteristic Curve: Spek 9-12 S501 Online



2.9 Test Information Function

With the Rasch measurement model, as with any measurement model following item response theory, one can use the item information function (Lord, 1980) to model the relationship between the ability measure (in logits) and the accuracy of the ability measure by item. The item information function indicates the amount of information we have about the ability estimate provided by the item, as a function of the ability level. The more information we have about the ability estimate, the more certain or confident we are about the ability estimate. If the amount of information is large, that means the student whose true ability is at that level is estimated with a higher degree of certainty, and all the estimates will be reasonably close to the true values. Conversely, if the amount of information is small, that means the student whose true ability is at that level is estimated with a lower degree of certainty and estimates will be further away from the true values. Mathematically, the amount of item information at a given ability level is the reciprocal of the variance of the ability estimate at the level for the item. In other words, item information value is the inverse squared of the standard errors of measurement of a given ability measure for the item. Therefore, item information is also said to provide information about the precision of the ability estimate along the ability continuum provided by the item.

The test information function (TIF) aggregates the item information functions across all the items on the test form or item pool. Since the item information value is the inverse squared of the standard errors of measurement of a given ability measure for the item, the TIF reflects the standard errors of measurement of a given ability level for the test. 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, certainty, or the amount of measurement precision the test provides at each ability level. The higher the curve, the more information the test provides at the ability level.

Since the TIF is the sum of all item characteristic functions on the test form (Lord, 1980), the TIF depends on the item information functions (Lord, 1980) of the items 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, the item response theory model used, and the values of the item parameters. With some exceptions, there is a general pattern to the shape of test information curves. Test information curves peak at the area where the test provides higher discrimination and better measurement as compared to other areas 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 is usually unimodal. The shape of test information curves for Writing and Speaking tests, which consist of polytomous tasks, are affected by the values of the item category parameters in addition to the factors mentioned earlier. Since polytomous tasks have more score categories than multiple-choice items and they measure a wider range of values on the proficiency scale, adjacent category boundaries are sometimes far apart as a result. In this situation, a test information curve will have a dip in the area between the adjacent category boundaries, indicating the loss of

information in this ability range. Therefore, the shape of a test information curve for ACCESS Writing and Speaking tests may not be unimodal and instead may have one or more peaks. This is consistent with other tests with polytomous items, such as the National Assessment of Educational Progress Writing assessment (Muraki, 1993).

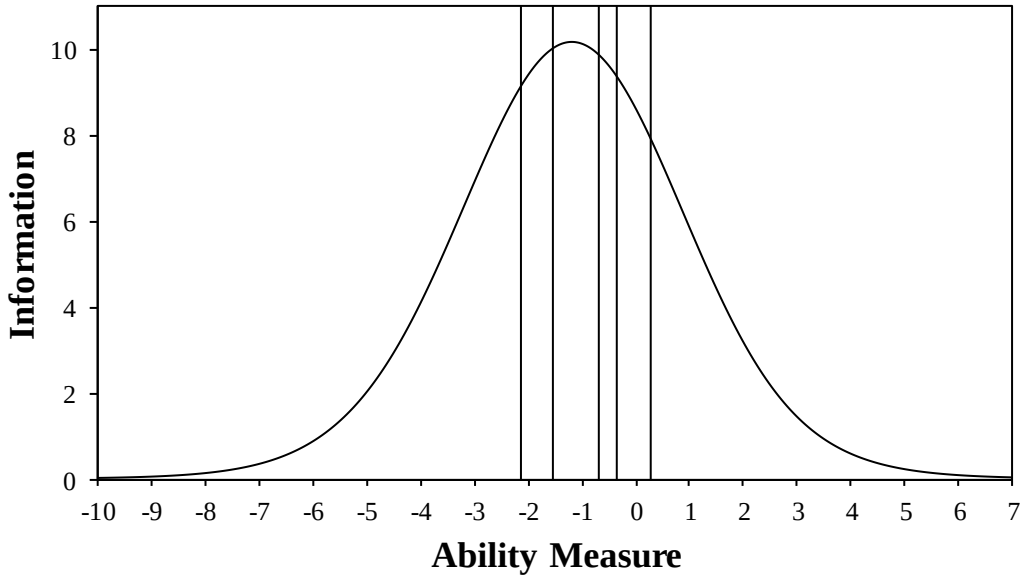
The figures in this section plot the TIF and show graphically the amount of information provided by the test across the continuum of student ability. Five vertical lines in the figure indicate the five ACCESS cut scores for the highest grade in the grade-level cluster for the test form, dividing the figure into six sections for each of the WIDA proficiency levels (1–6) for the domain being tested. The ACCESS cut score lines are presented along with the TIF to facilitate the interpretation of the test information curves. The test information curve and the corresponding ACCESS cut score lines are both expressed on the ACCESS logit scale. Note that for Speaking, in Tier Pre-A, all scores fall in the PL 1.0 range, so for some graphs there are no vertical lines expressing the cuts between proficiency levels.

In addition to the TIF graphs by tier, for Writing and Speaking, we provide plots of the TIFs across tiers, by grade cluster, on the same graph. It is informative to compare the ability ranges where the curves are peaked (where the best measurement information is provided) across tiers. For example, the test information curve across tiers for Writing Grade 1 shows that the Writing Grade 1 Tier A form provides more information just below the PL 2 cut, and also just below the PL 4 cut. The Writing Grade 1 Tier B/C form provides more information just above the PL 2 cut, and just above the PL 4 cut. The plot also shows that the Writing Grade 1 Tier A form provides more information at the lowest ability range (ability measure of 5.0 or lower) while the Writing Grade 1 Tier B/C form provides more information than the Grade 1 Tier A form for the rest of the ability range, especially at the higher ability range.

2.9.1 **Listening**

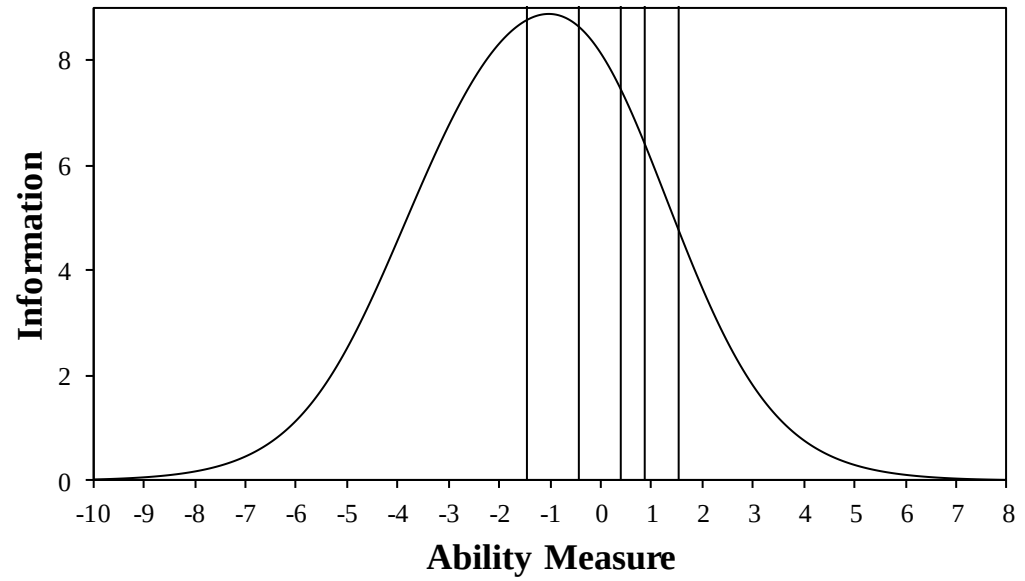
2.9.1.1 *Grade 1*

Figure 2.9.1.1
Test Information Function: List 1 S501 Online



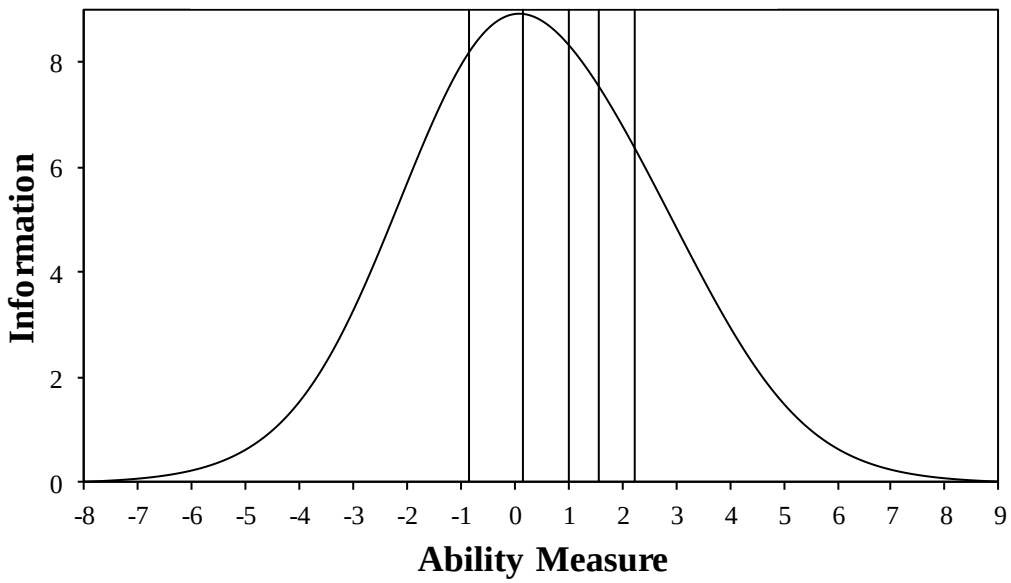
2.9.1.2 *Grades 2–3*

Figure 2.9.1.2
Test Information Function: List 2-3 S501 Online



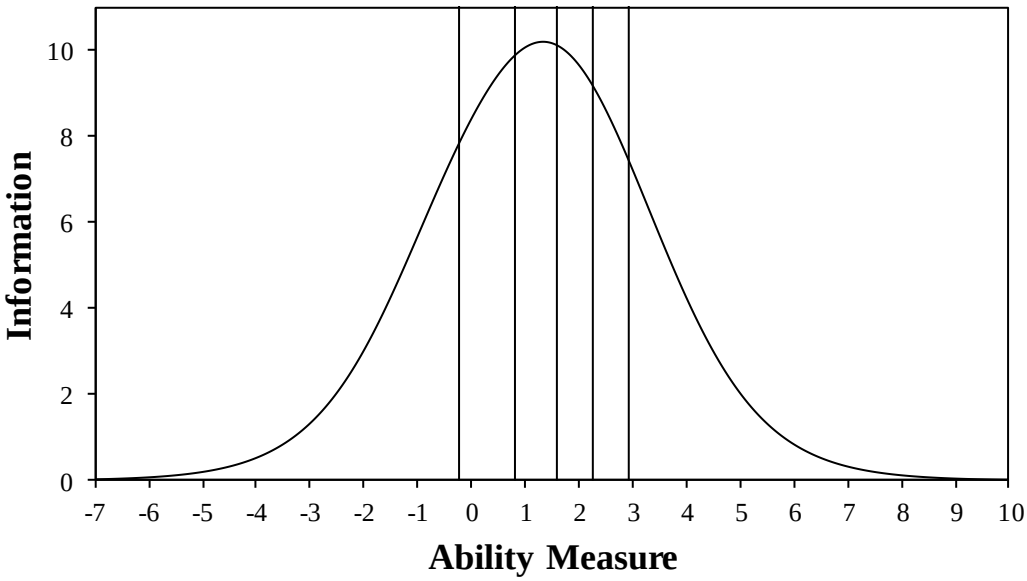
2.9.1.3 *Grades 4–5*

Figure 2.9.1.3
Test Information Function: List 4-5 S501 Online



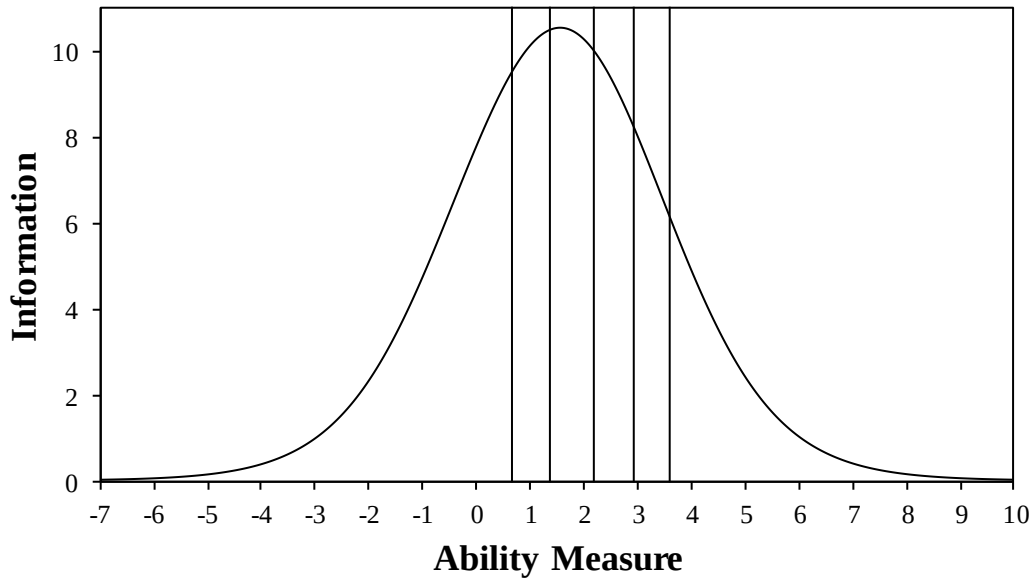
2.9.1.4 *Grades 6–8*

Figure 2.9.1.4
Test Information Function: List 6-8 S501 Online



2.9.1.5 *Grades 9–12*

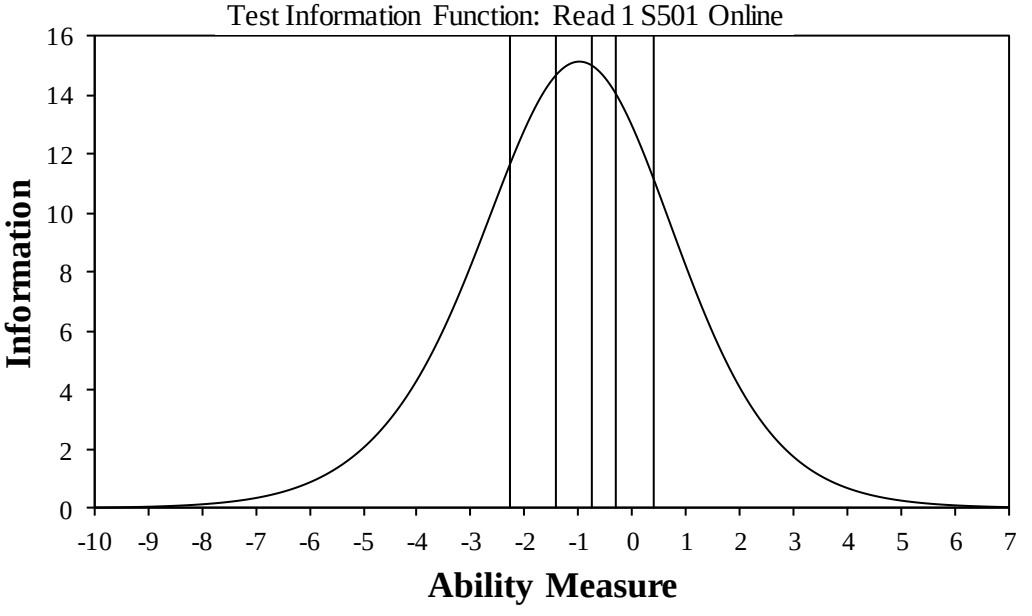
Figure 2.9.1.5
Test Information Function: List 9-12 S501 Online



2.9.2 **Reading**

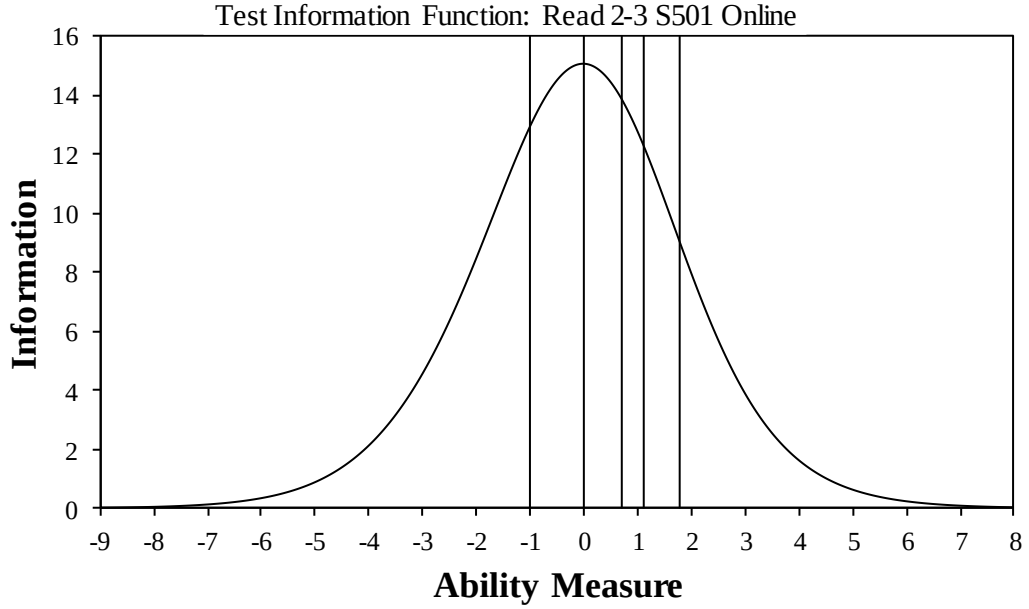
2.9.2.1 *Grade 1*

Figure 2.9.2.1
Test Information Function: Read 1 S501 Online



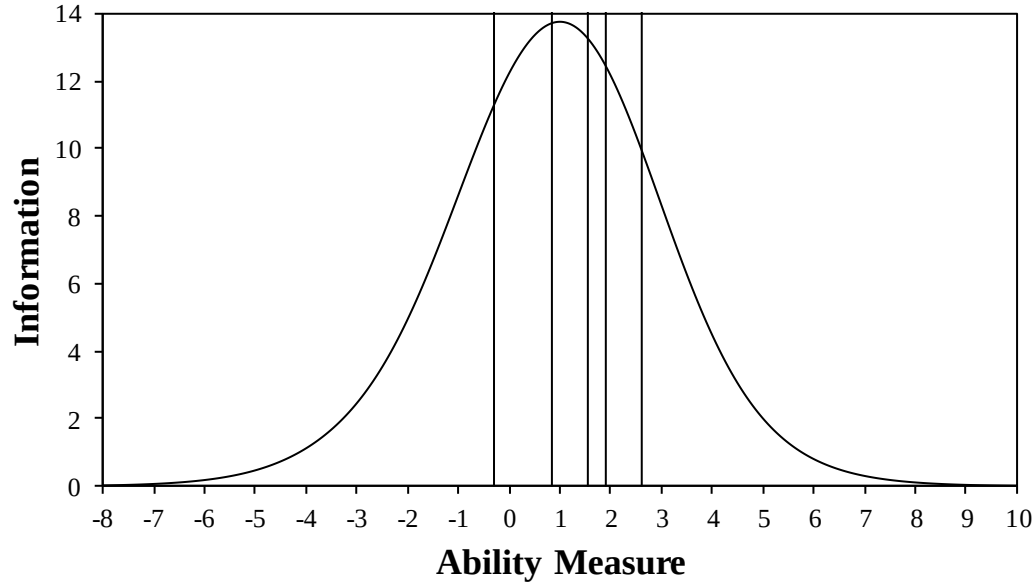
2.9.2.2 *Grades 2–3*

Figure 2.9.2.2
Test Information Function: Read 2-3 S501 Online



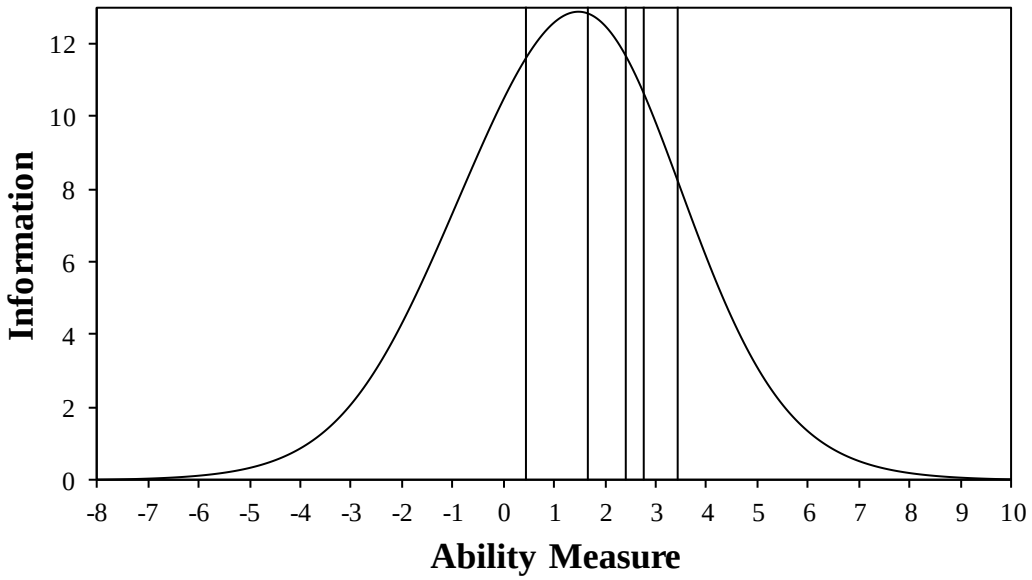
2.9.2.3 *Grades 4–5*

Figure 2.9.2.3
Test Information Function: Read 4-5 S501 Online



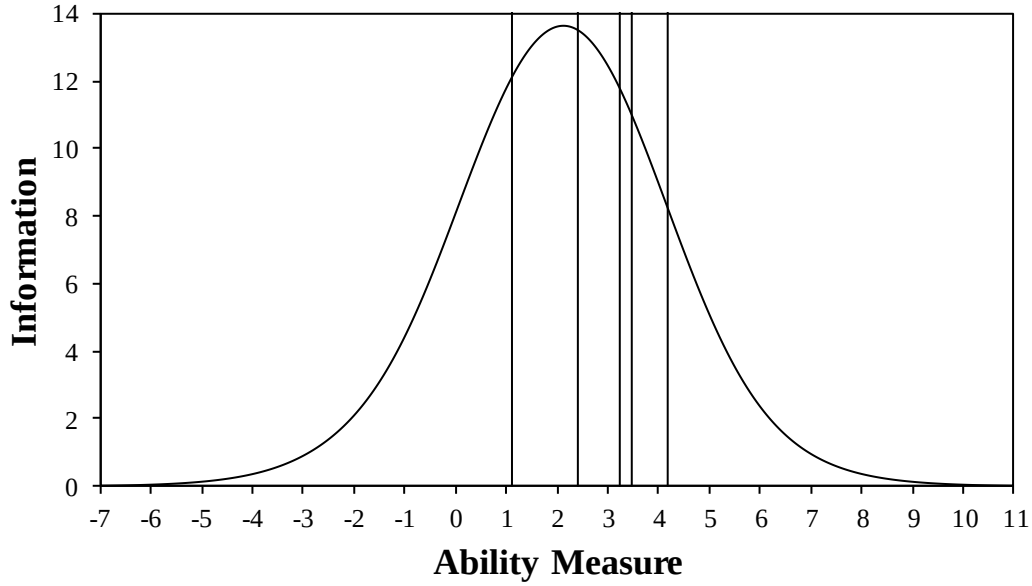
2.9.2.4 *Grades 6–8*

Figure 2.9.2.4
Test Information Function: Read 6-8 S501 Online



2.9.2.5 *Grades 9–12*

Figure 2.9.2.5
Test Information Function: Read 9-12 S501 Online



2.9.3 **Writing**

2.9.3.1 *Grade 1*

Figure 2.9.3.1.1
Test Information Function: Writ 1 A S501 Online

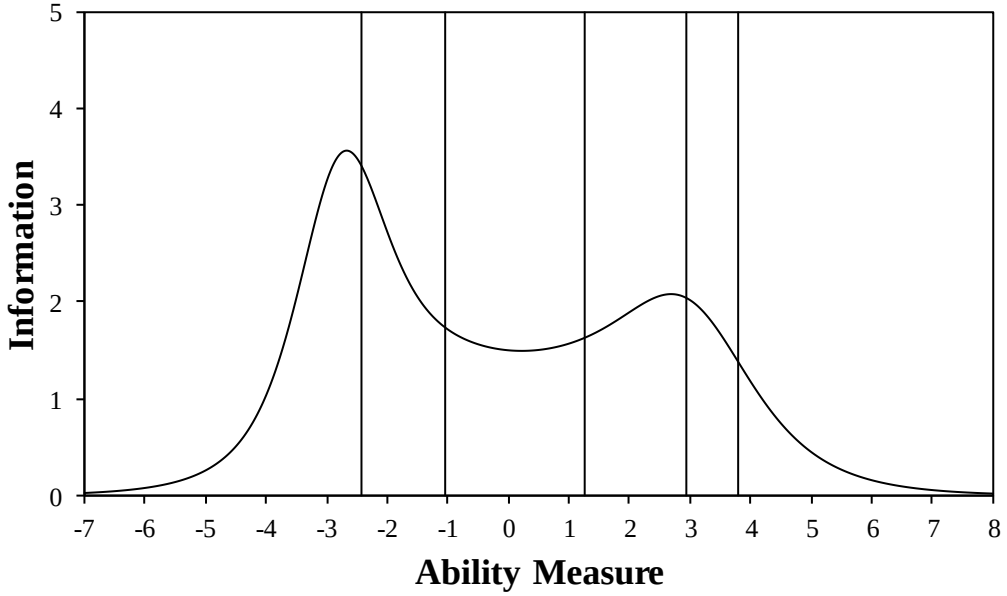


Figure 2.9.3.1.2
Test Information Function: Writ 1 B/C S501 Online

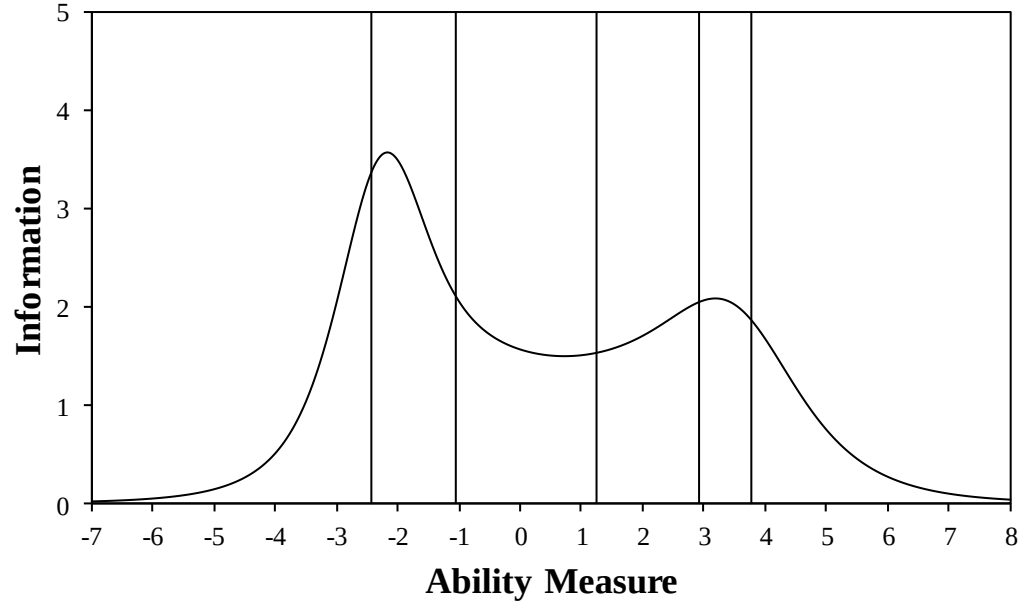
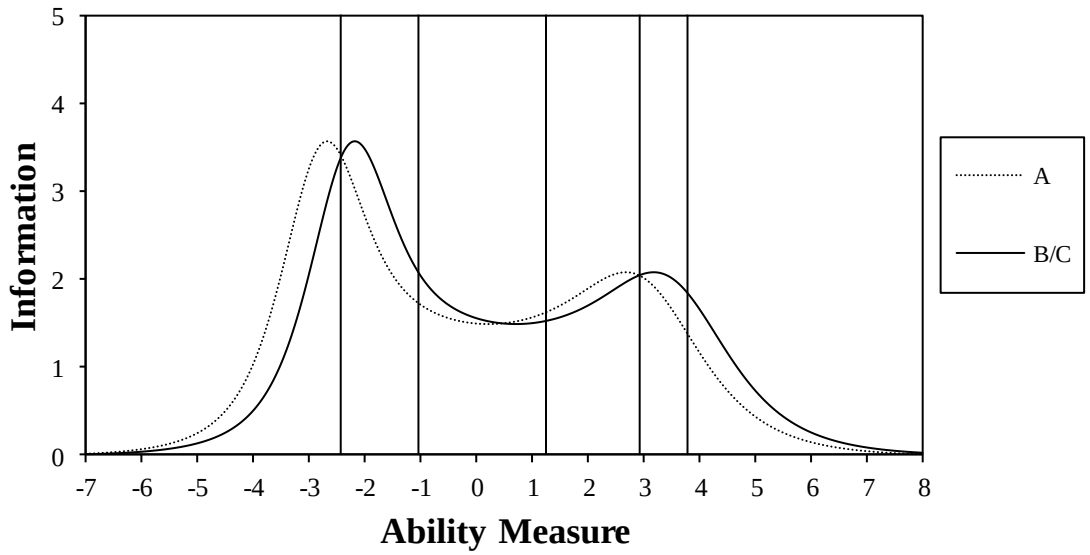


Figure 2.9.3.1.3
Test Information Function: Writ 1 S501 Online



2.9.3.2 *Grades 2–3*

Figure 2.9.3.2.1
Test Information Function: Writ 2-3 A S501 Online

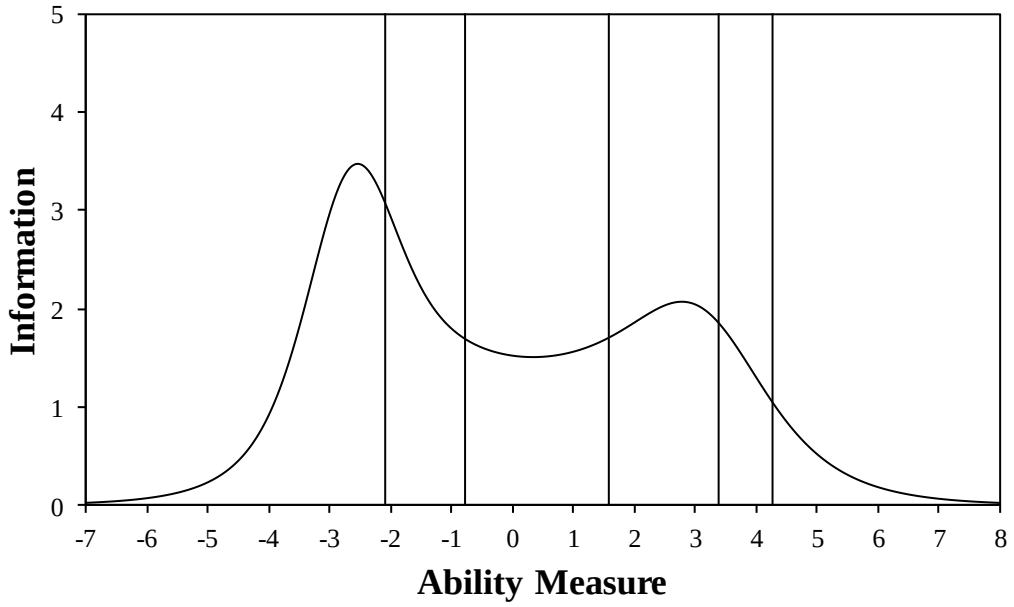


Figure 2.9.3.2.2
Test Information Function: Writ 2-3 B/C S501 Online

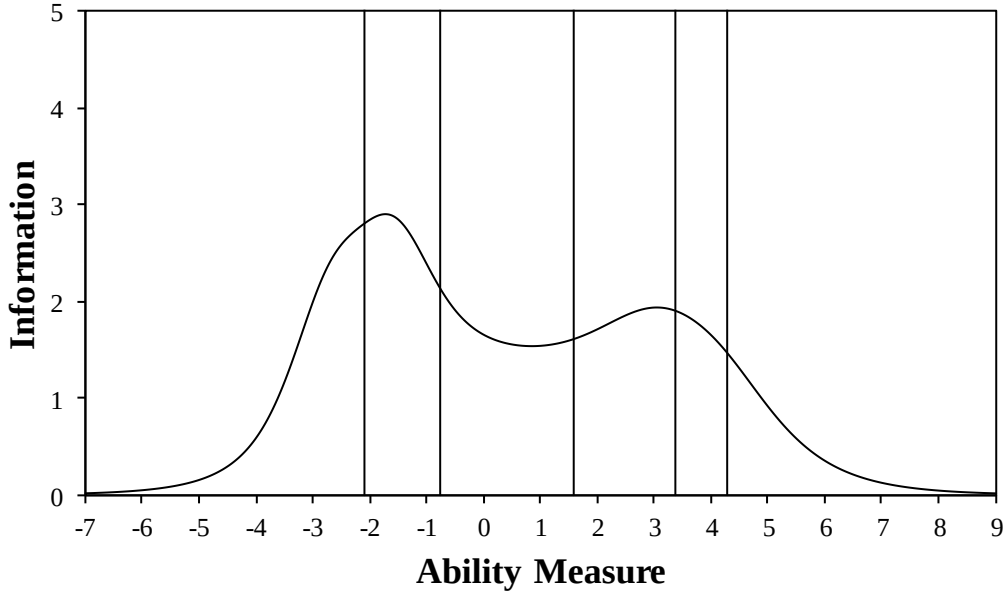
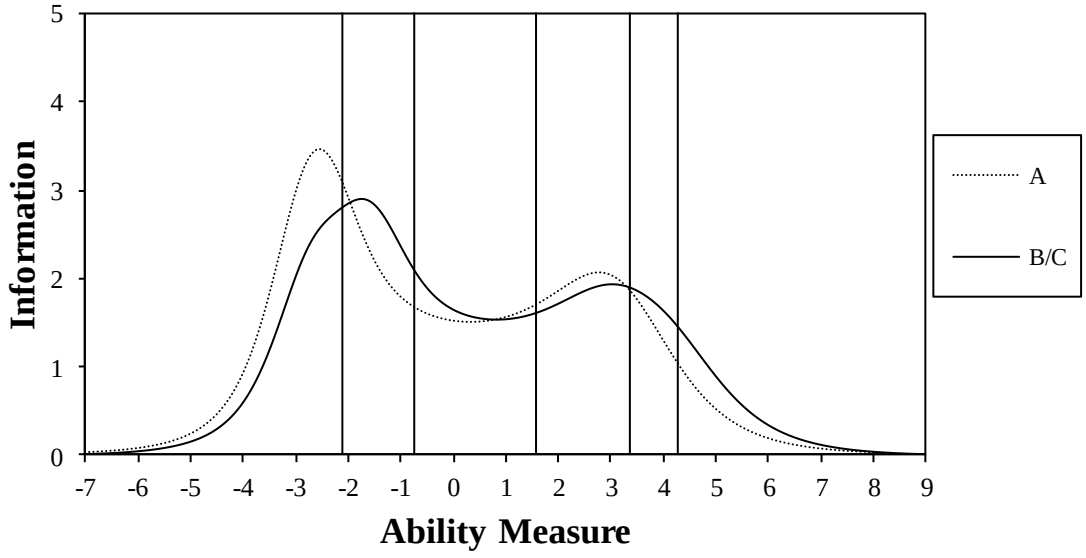


Figure 2.9.3.2.3
Test Information Function: Writ 2-3 S501 Online



2.9.3.3 *Grades 4–5*

Figure 2.9.3.3.1
Test Information Function: Writ 4-5 A S501 Online

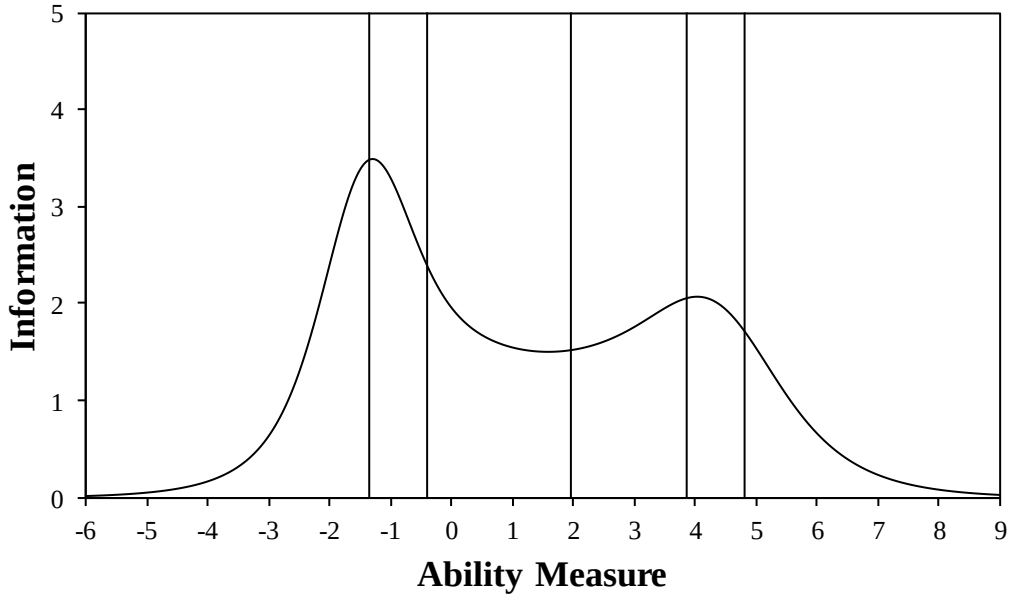


Figure 2.9.3.3.2
Test Information Function: Writ 4-5 B/C S501 Online

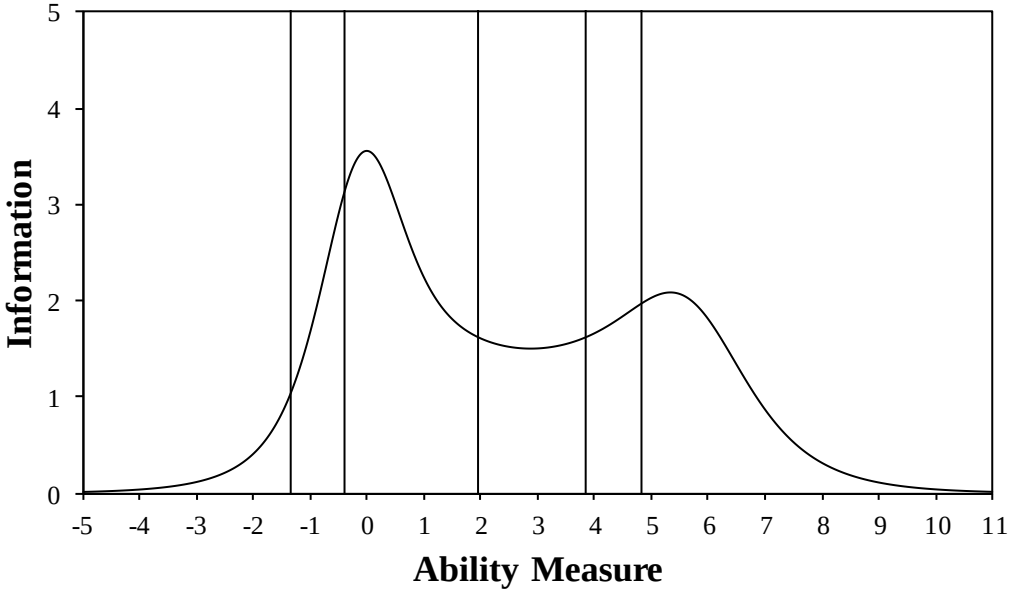
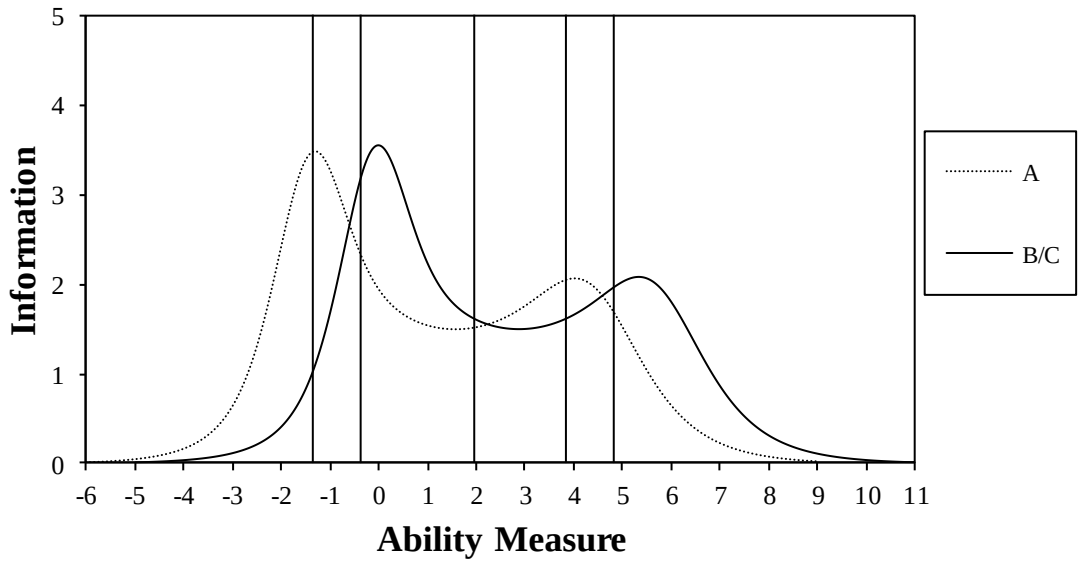


Figure 2.9.3.3.3
Test Information Function: Writ 4-5 S501 Online



2.9.3.4 *Grades 6–8*

Figure 2.9.3.4.1
Test Information Function: Writ 6-8 A S501 Online

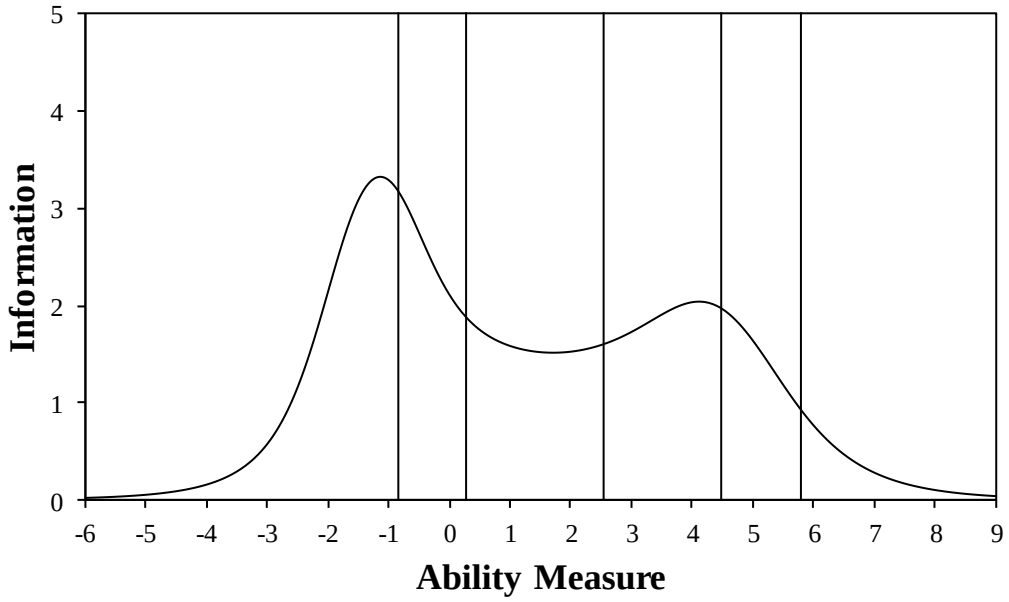


Figure 2.9.3.4.2
Test Information Function: Writ 6-8 B/C S501 Online

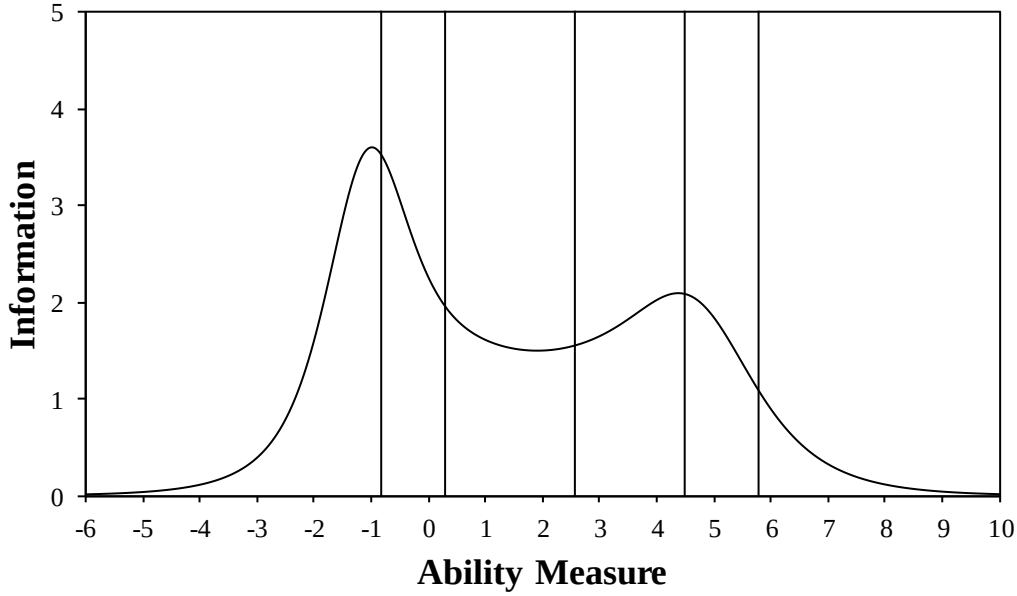
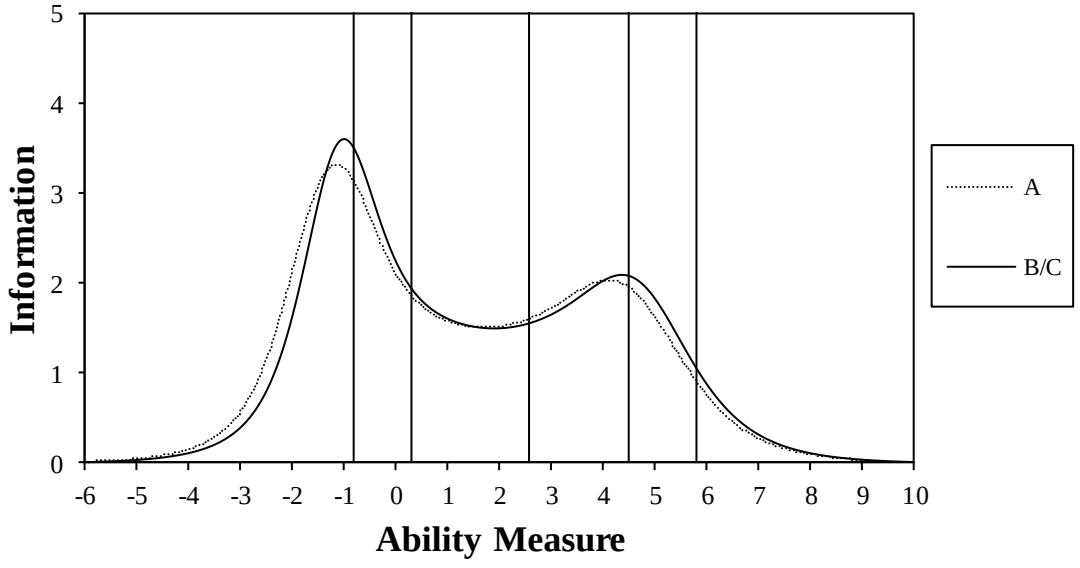


Figure 2.9.3.4.3
Test Information Function: Writ 6-8 S501 Online



2.9.3.5 *Grades 9–12*

Figure 2.9.3.5.1
Test Information Function: Writ 9-12 A S501 Online

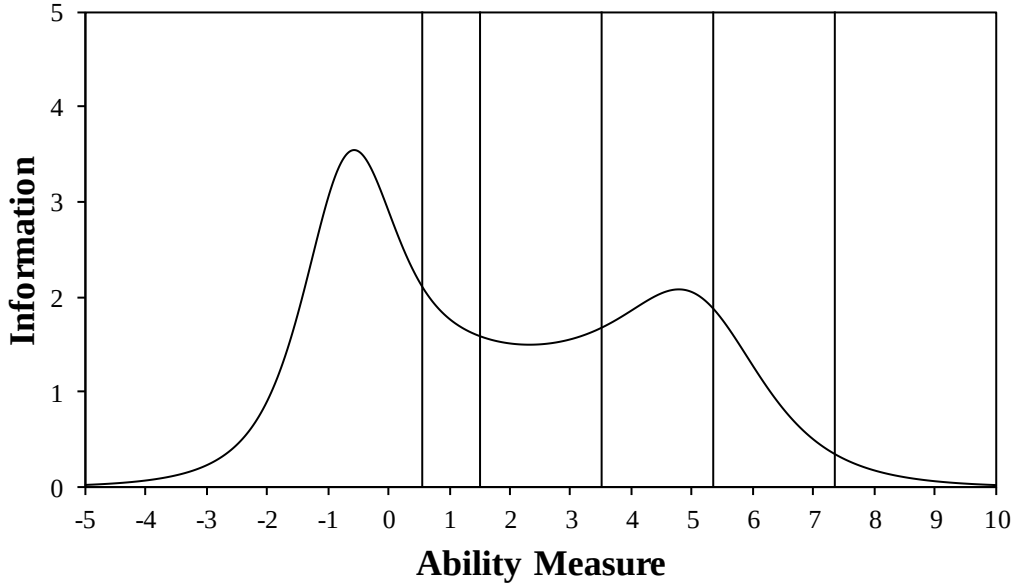


Figure 2.9.3.5.2
Test Information Function: Writ 9-12 B/C S501 Online

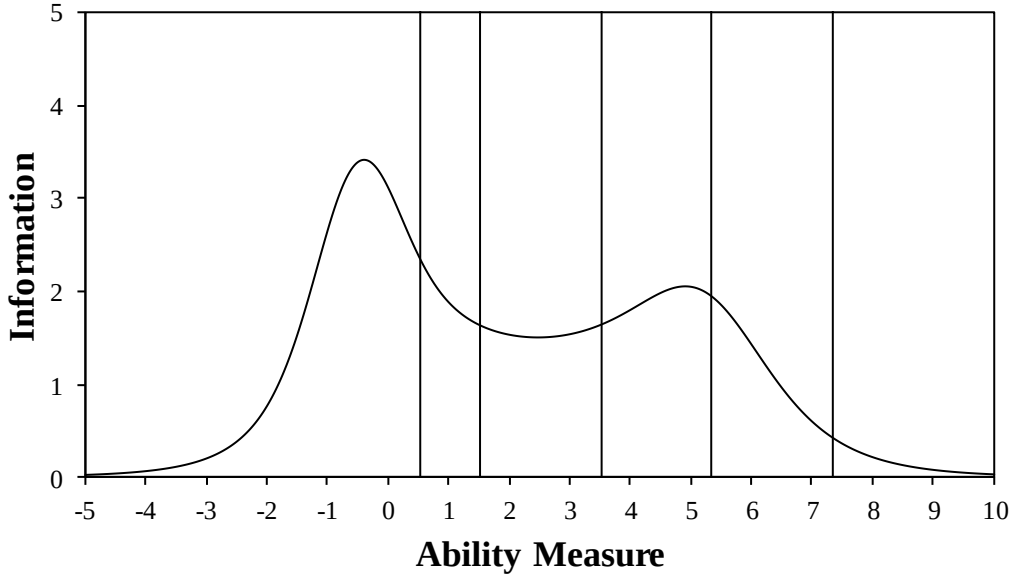
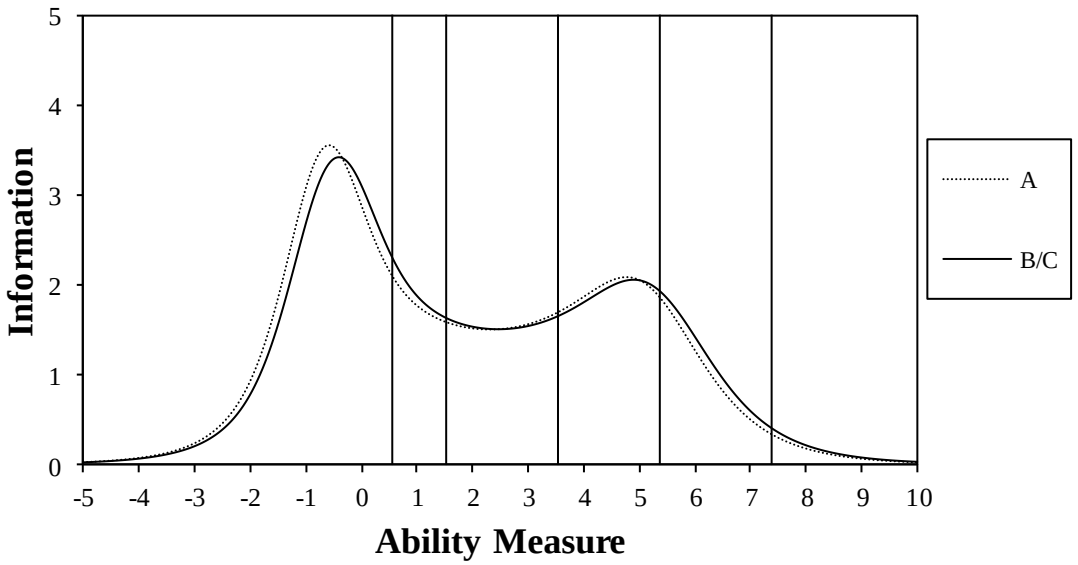


Figure 2.9.3.5.3
Test Information Function: Writ 9-12 S501 Online



2.9.4 **Speaking**

2.9.4.1 *Grade 1*

Figure 2.9.4.1.1
Test Information Function: Spek 1 Pre-A S501 Online

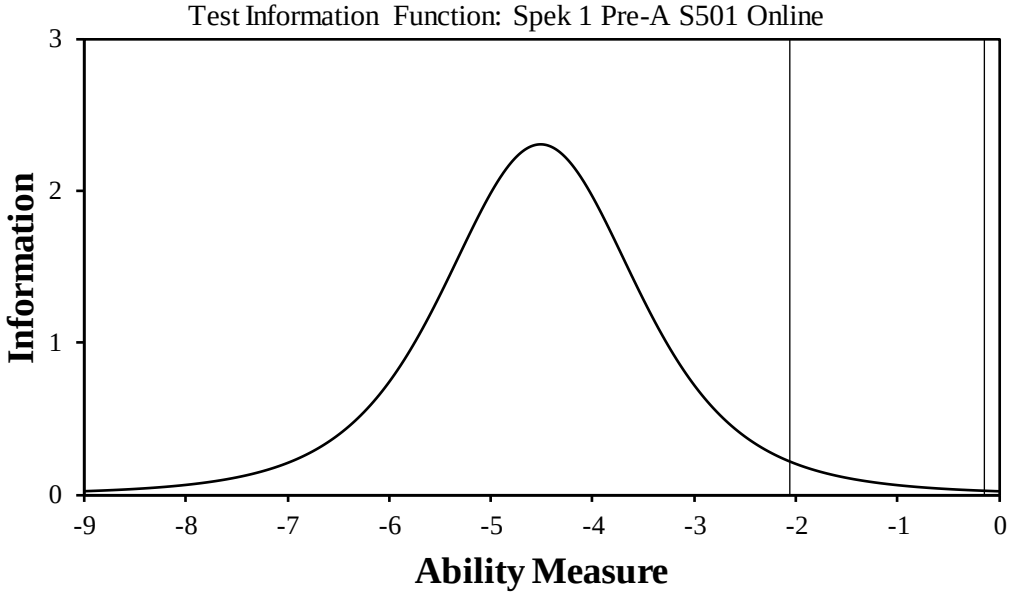


Figure 2.9.4.1.2
Test Information Function: Spek 1 A S501 Online

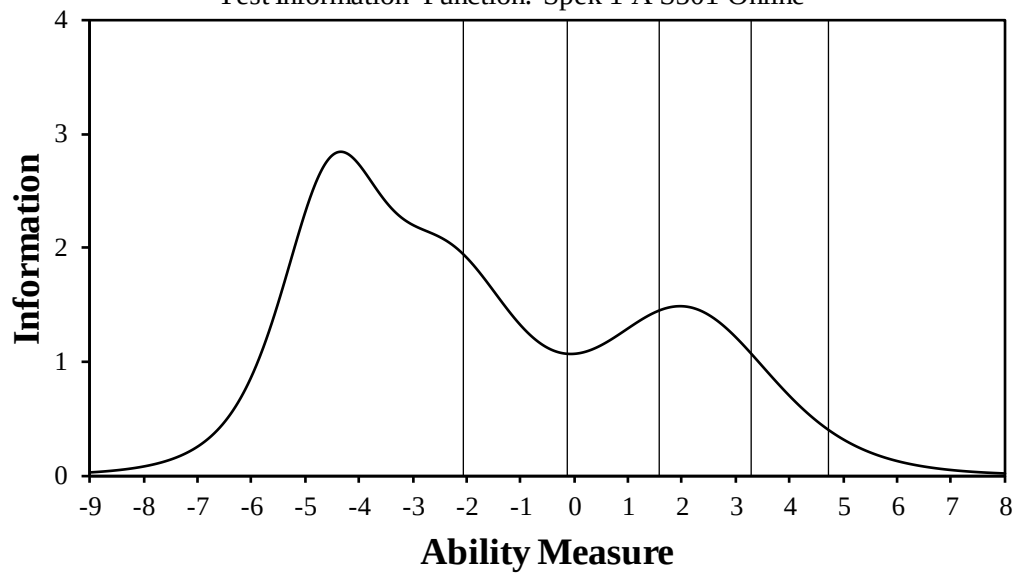


Figure 2.9.4.1.3
Test Information Function: Spek 1 B/C S501 Online

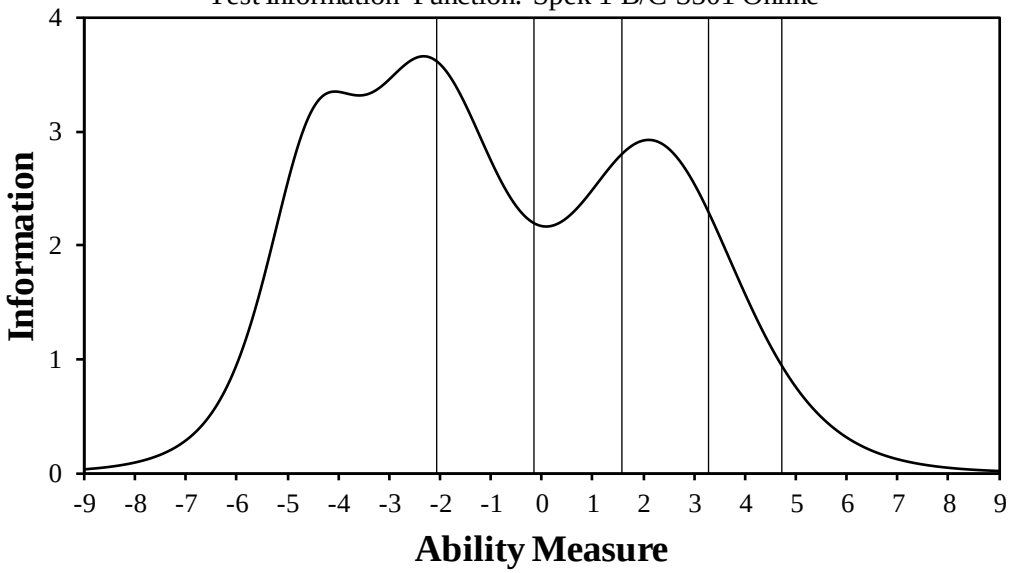
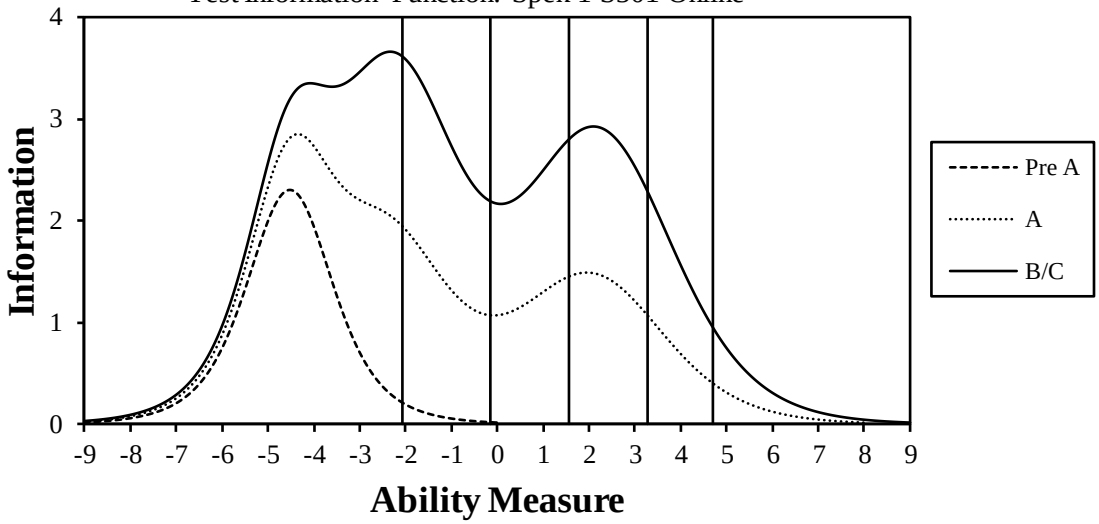


Figure 2.9.4.1.4
Test Information Function: Spek 1 S501 Online



2.9.4.2 *Grades 2–3*

Figure 2.9.4.2.1
Test Information Function: Spek 2-3 Pre-A S501 Online

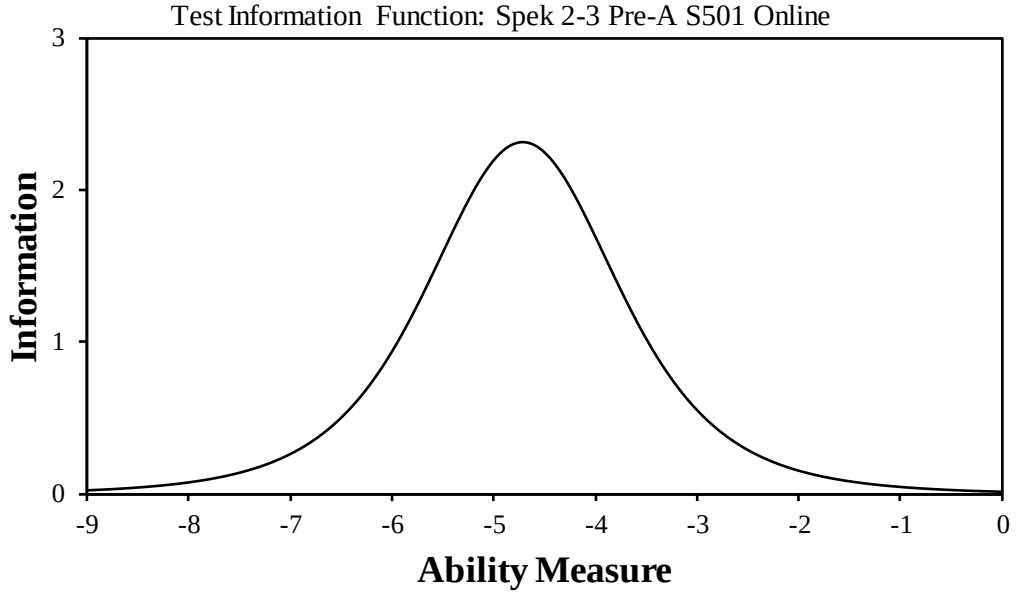


Figure 2.9.4.2.2
Test Information Function: Spek 2-3 A S501 Online

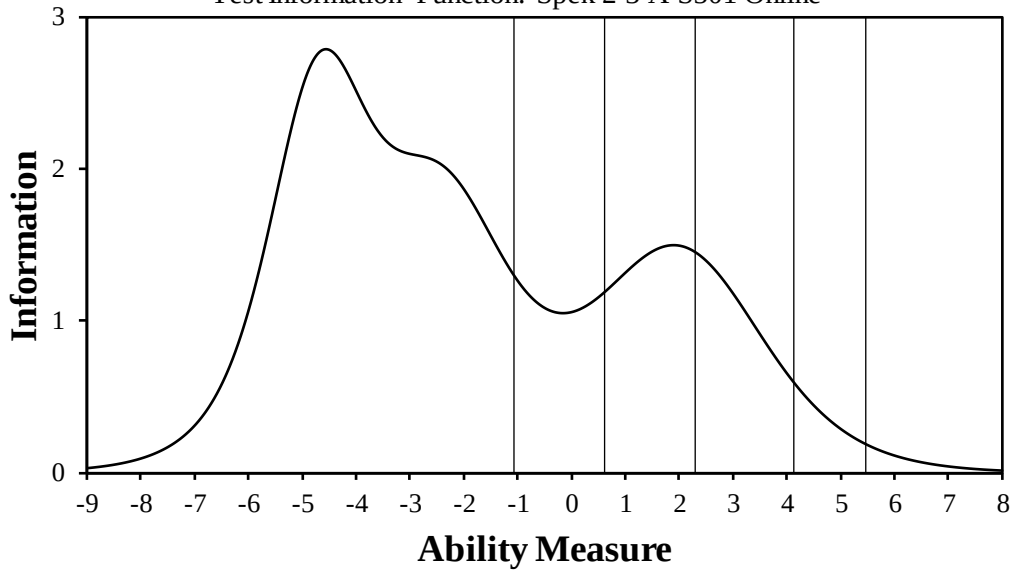


Figure 2.9.4.2.3
Test Information Function: Spek 2-3 B/C S501 Online

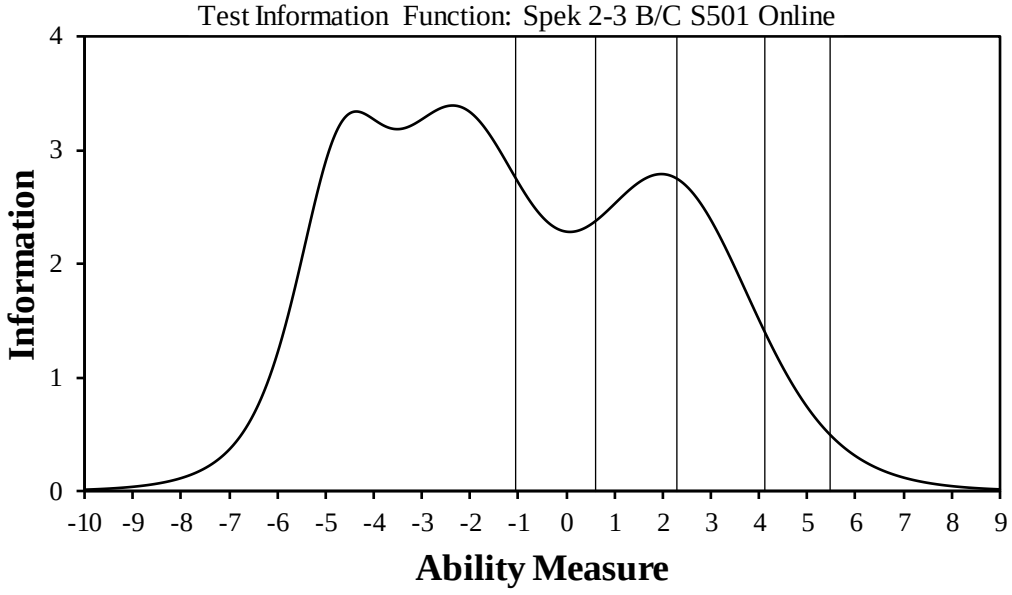
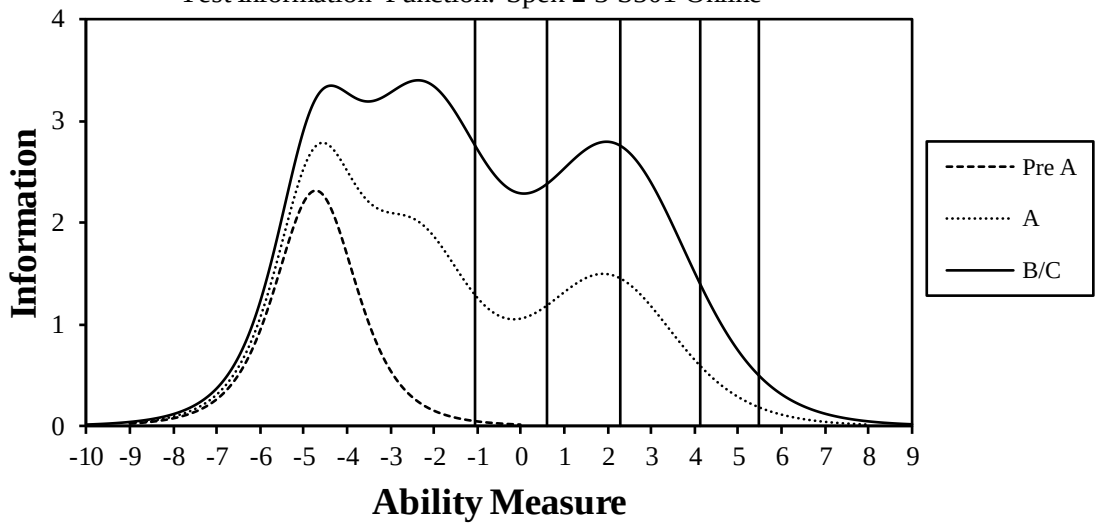


Figure 2.9.4.2.4
Test Information Function: Spek 2-3 S501 Online



2.9.4.3 *Grades 4–5*

Figure 2.9.4.3.1
Test Information Function: Spek 4-5 Pre-A S501 Online

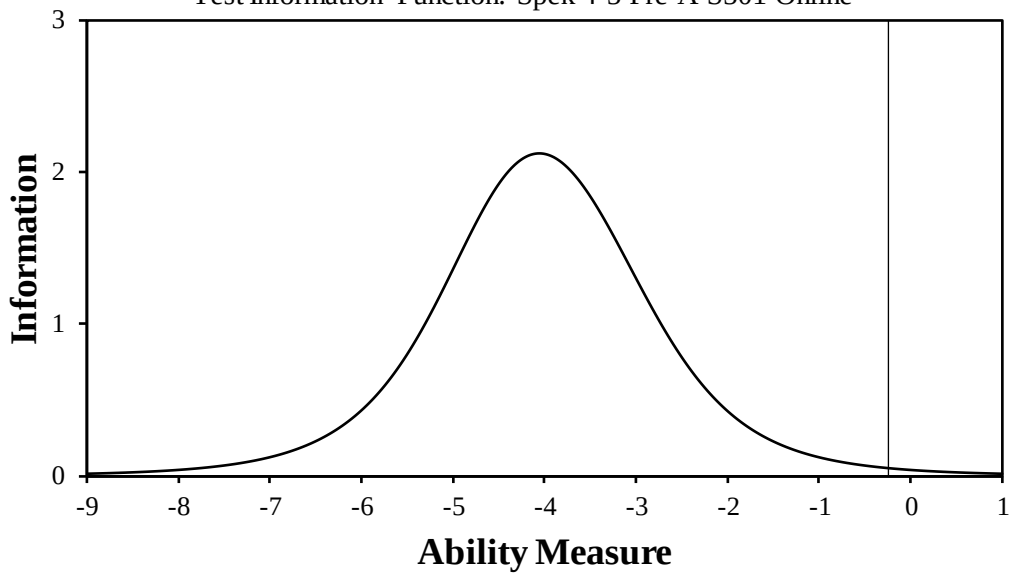


Figure 2.9.4.3.2
Test Information Function: Spek 4-5 A S501 Online

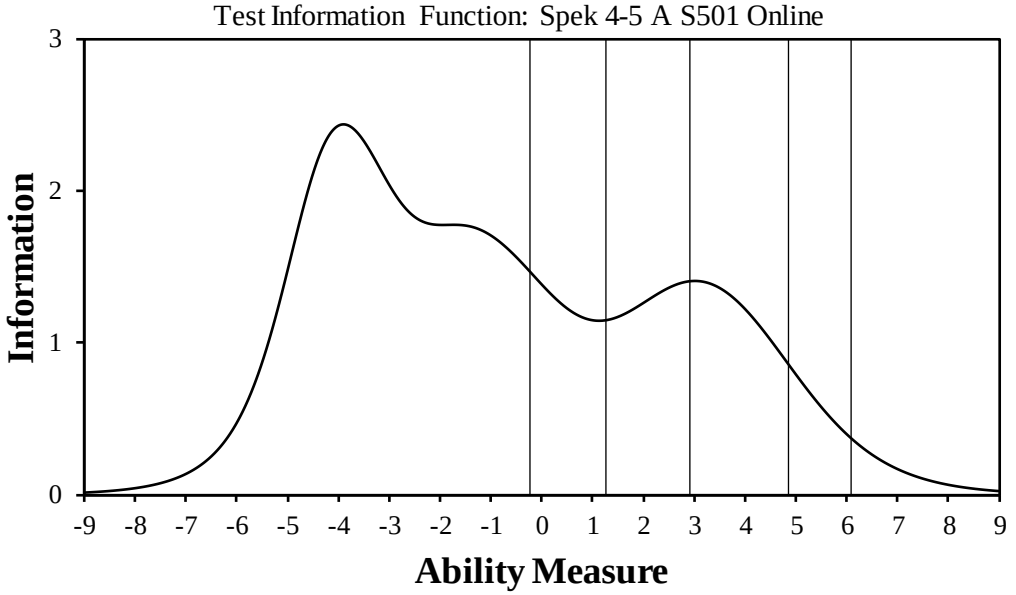


Figure 2.9.4.3.3
Test Information Function: Spek 4-5 B/C S501 Online

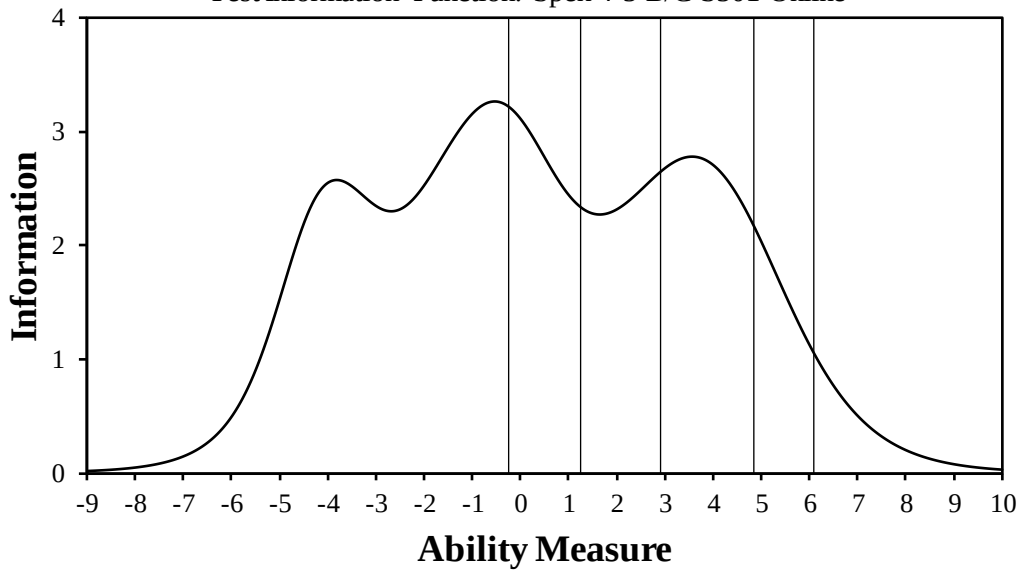
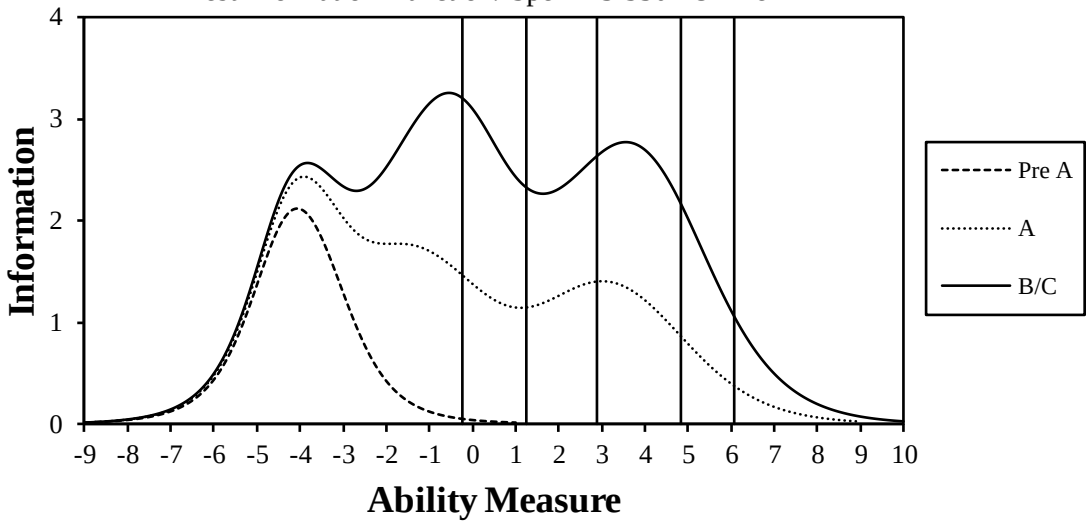


Figure 2.9.4.3.4
Test Information Function: Spek 4-5 S501 Online



2.9.4.4 *Grades 6–8*

Figure 2.9.4.4.1
Test Information Function: Spek 6-8 Pre-A S501 Online

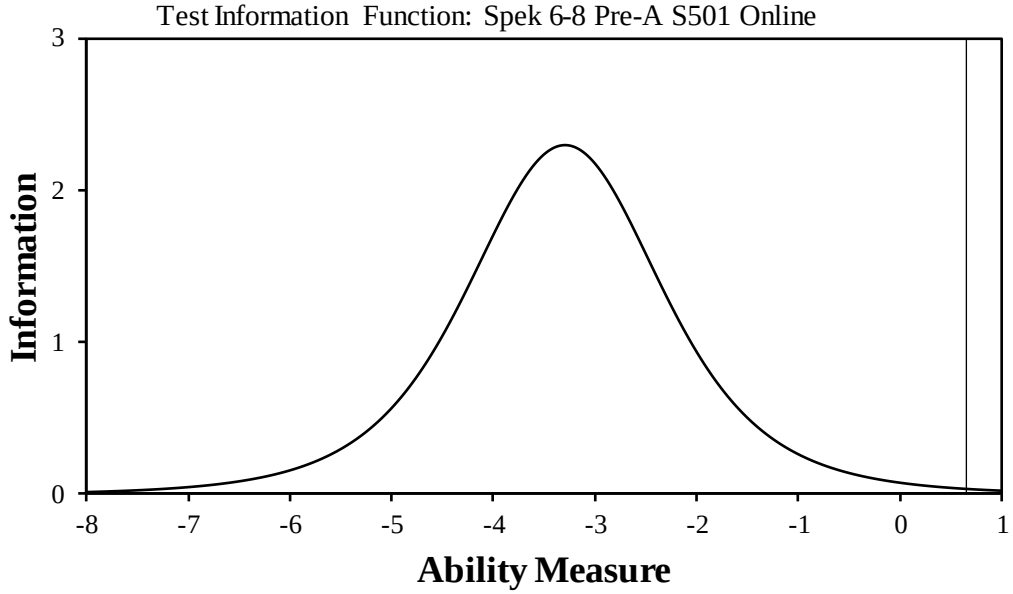


Figure 2.9.4.4.2
Test Information Function: Spek 6-8 A S501 Online

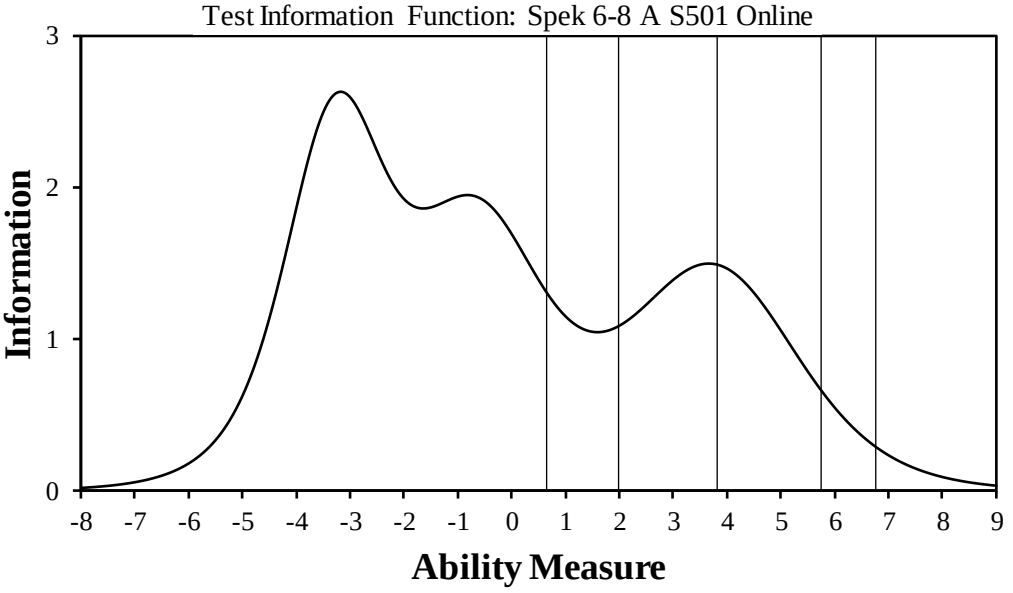


Figure 2.9.4.4.3
Test Information Function: Spek 6-8 B/C S501 Online

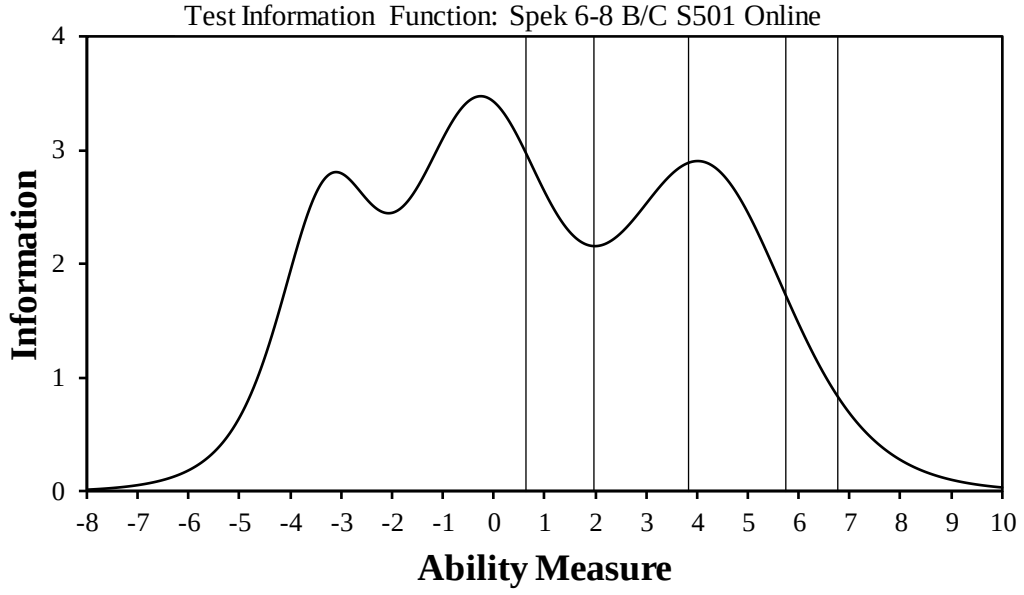
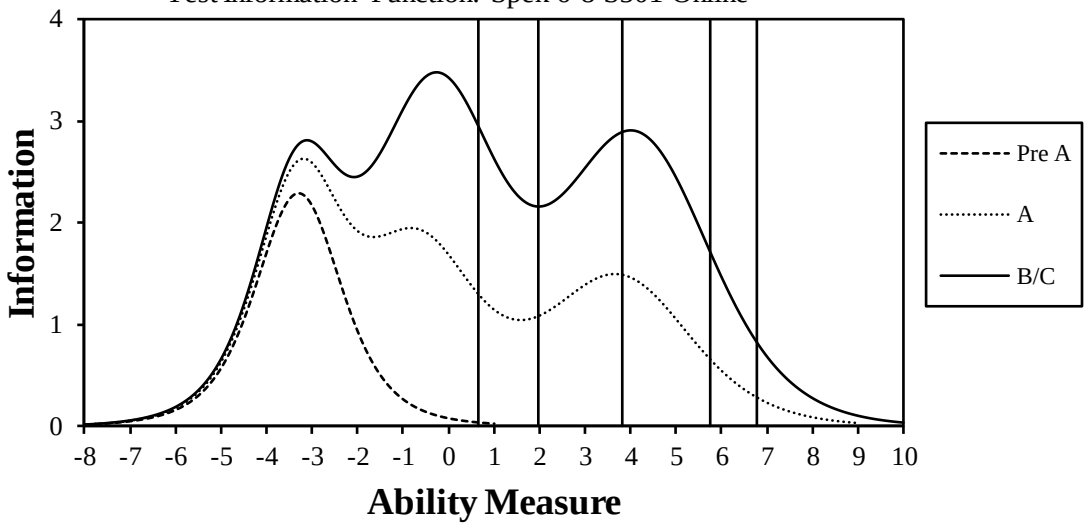


Figure 2.9.4.4.4
Test Information Function: Spek 6-8 S501 Online



2.9.4.5 *Grades 9–12*

Figure 2.9.4.5.1
Test Information Function: Spek 9-12 Pre-A S501 Online

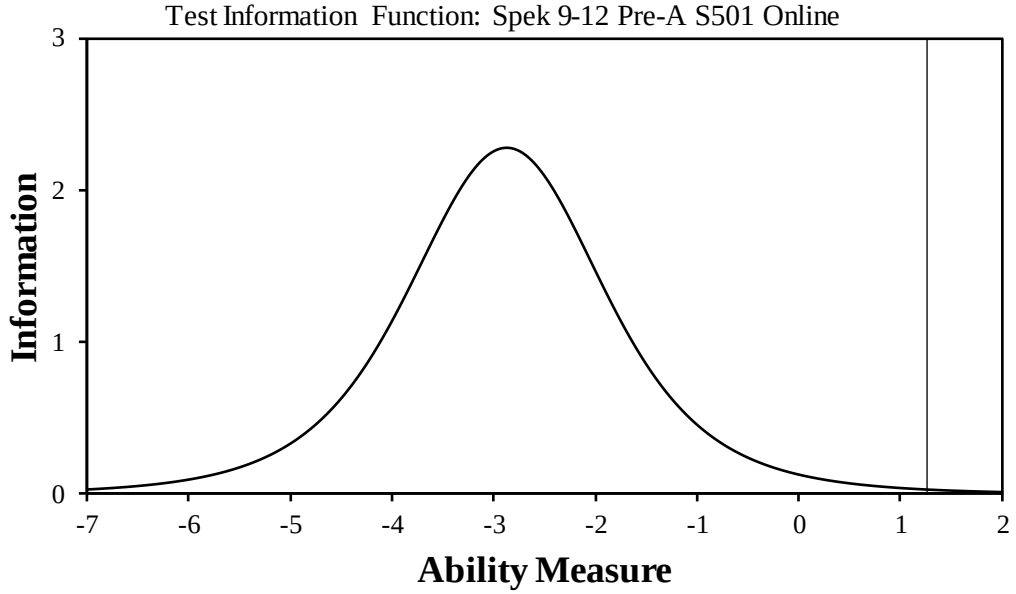


Figure 2.9.4.5.2
Test Information Function: Spek 9-12 A S501 Online

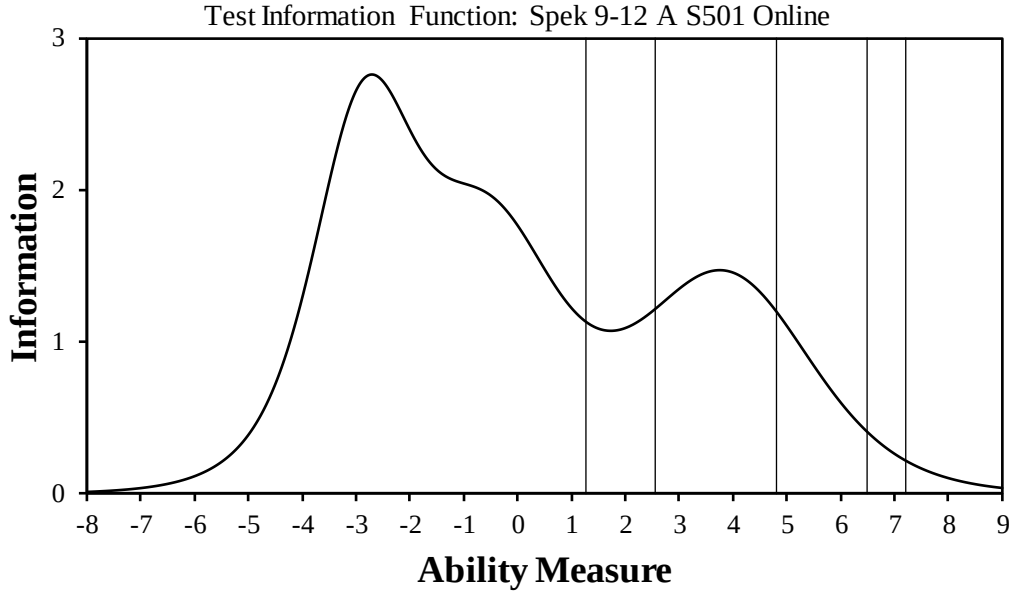


Figure 2.9.4.5.3
Test Information Function: Spek 9-12 B/C S501 Online

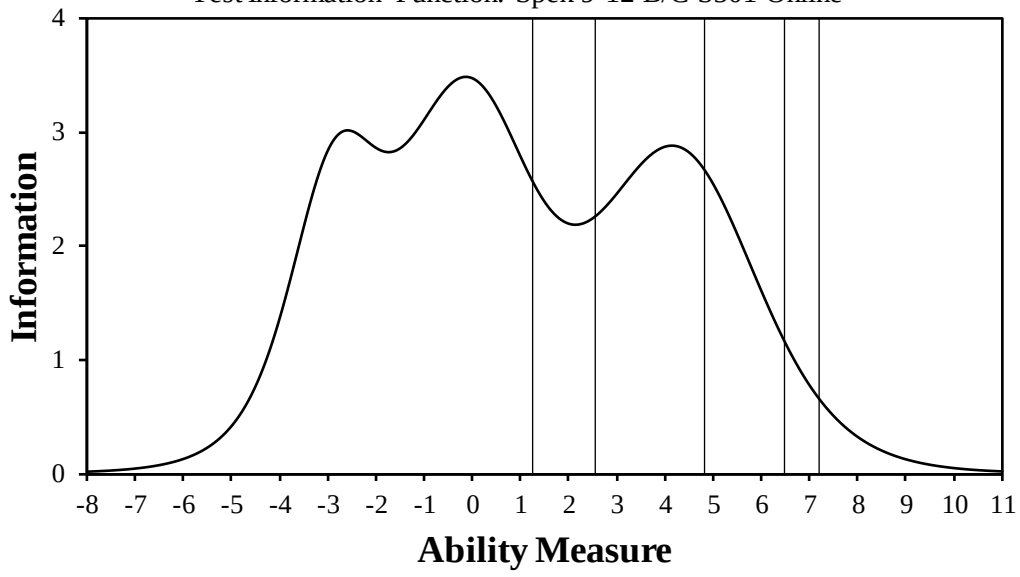
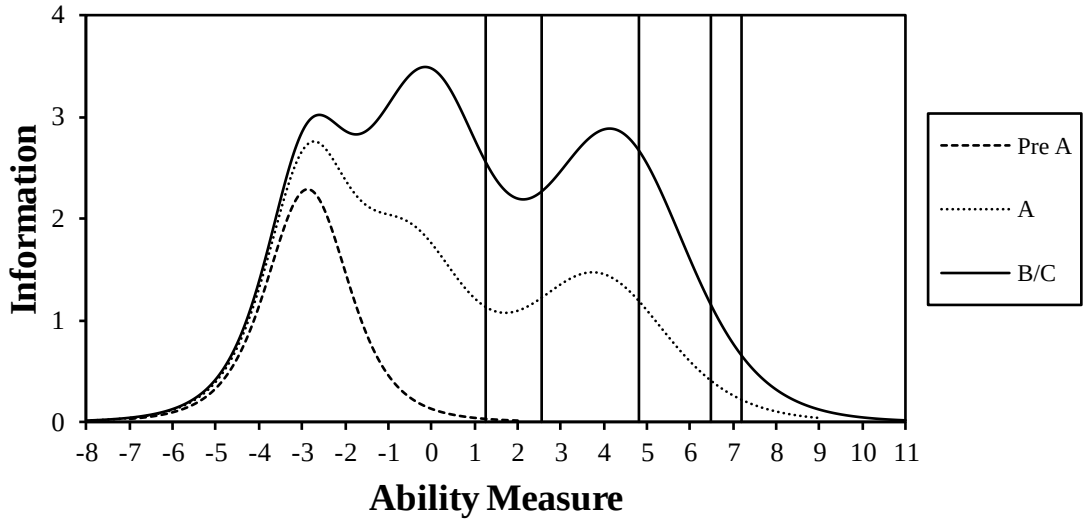


Figure 2.9.4.5.4
Test Information Function: Spek 9-12 S501 Online



3 Analyses of Composite Scores

Four composite scores are calculated for ACCESS Online: Oral Language, Literacy, Comprehension, and Overall. Composite scores are calculated 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

This weighting resulted from a policy decision by the WIDA Board before the first operational administration of ACCESS, 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. Scale scores are plotted on the horizontal axis, grouped into units of five scale score points (e.g., 100–104, 105–109, 110–114, etc.). The number of students with scale scores falling into each range is plotted 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**

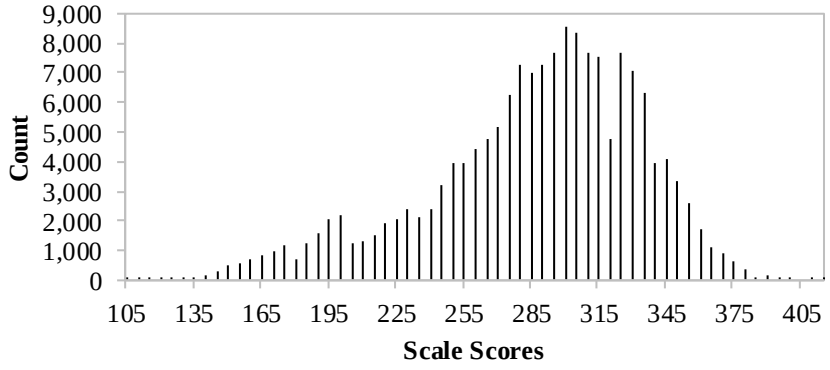
3.1.1.1 *Grade 1*

Table 3.1.1.1

Scale Score Descriptive Statistics: Oral 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	165,935	107	417	287.90	47.73
Total	165,935	107	417	287.90	47.73

Figure 3.1.1.1
Scale Scores: Oral 1 S501 Online



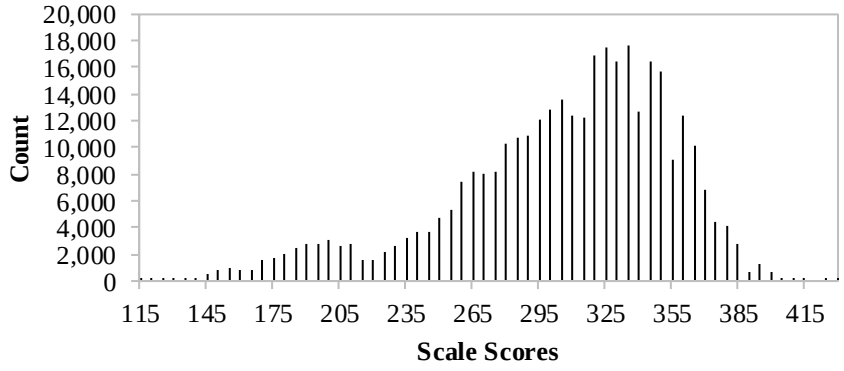
3.1.1.2 *Grades 2–3*

Table 3.1.1.2

Scale Score Descriptive Statistics: Oral 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	173,381	115	434	296.88	48.62
3	173,473	115	434	316.95	50.15
Total	346,854	115	434	306.92	50.40

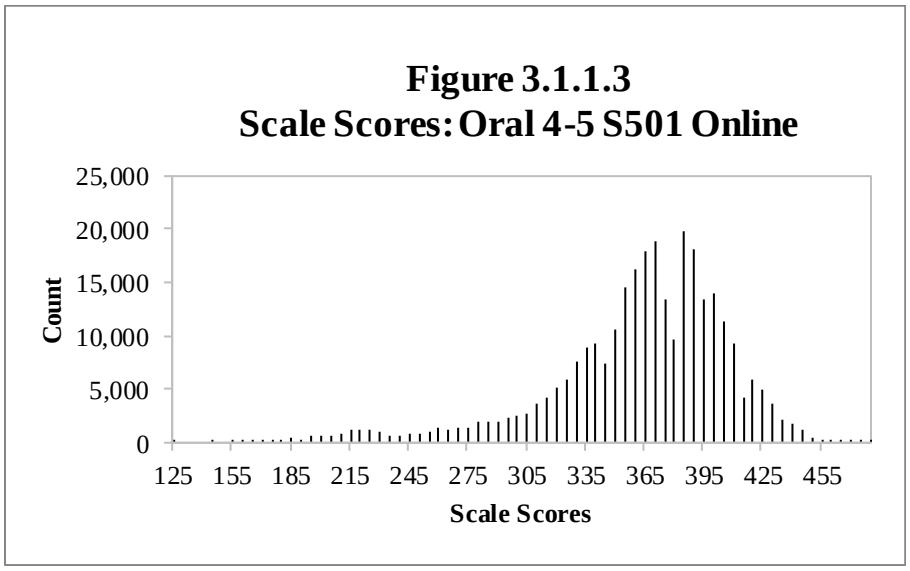
Figure 3.1.1.2
Scale Scores: Oral 2-3 S501 Online



3.1.1.3 *Grades 4–5*

Table 3.1.1.3
Scale Score Descriptive Statistics: Oral 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	163,102	127	483	362.52	44.31
5	132,035	145	483	366.36	47.80
Total	295,137	127	483	364.24	45.95



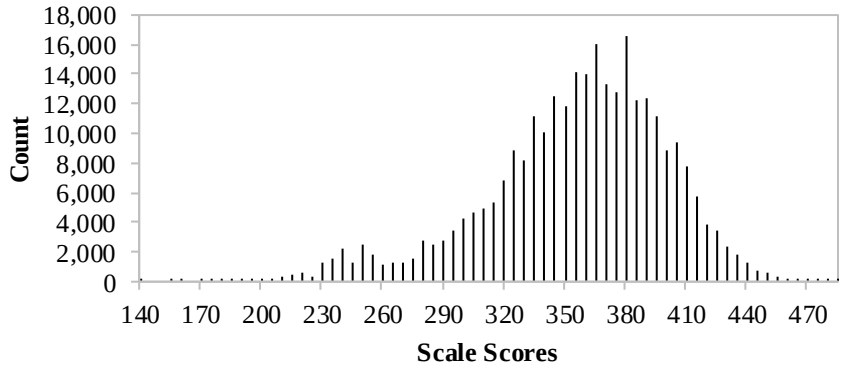
3.1.1.4 *Grades 6–8*

Table 3.1.1.4

Scale Score Descriptive Statistics: Oral 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	106,490	140	485	355.60	40.22
7	95,689	140	485	357.28	45.10
8	84,373	164	485	361.17	48.71
Total	286,552	140	485	357.80	44.54

Figure 3.1.1.4
Scale Scores: Oral 6-8 S501 Online



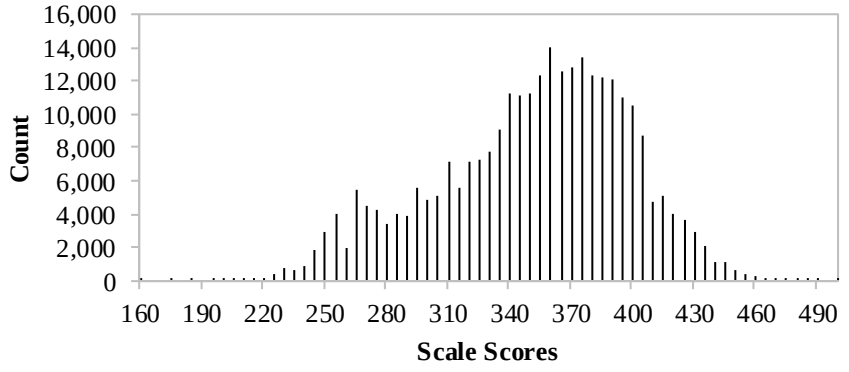
3.1.1.5 *Grades 9–12*

Table 3.1.1.5

Scale Score Descriptive Statistics: Oral 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	96,040	164	487	343.87	47.80
10	76,119	196	501	353.79	46.70
11	63,123	185	501	360.23	45.61
12	54,990	178	501	360.97	45.84
Total	290,272	164	501	353.27	47.22

Figure 3.1.1.5
Scale Scores: Oral 9-12 S501 Online

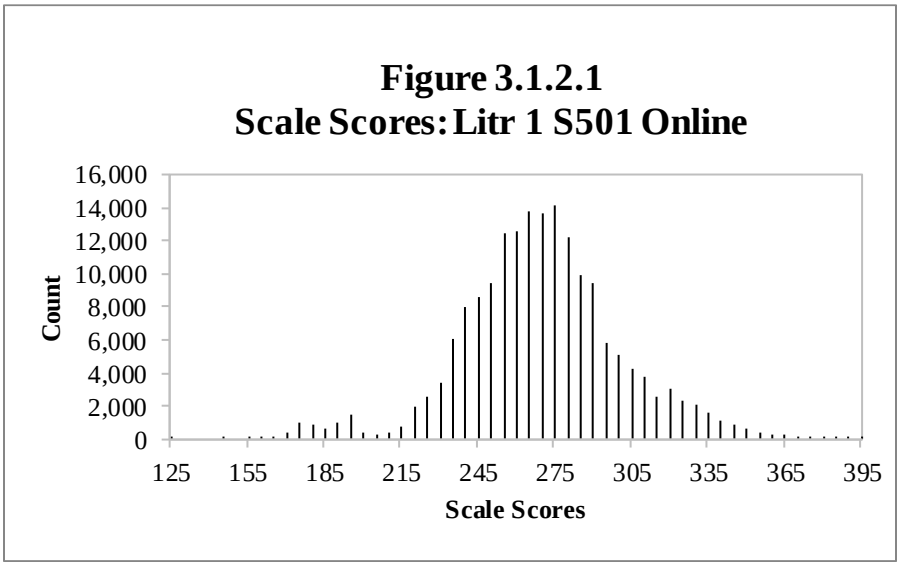


3.1.2 **Literacy**

3.1.2.1 *Grade 1*

Table 3.1.2.1
Scale Score Descriptive Statistics: Litr 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	179,697	126	399	271.26	31.36
Total	179,697	126	399	271.26	31.36

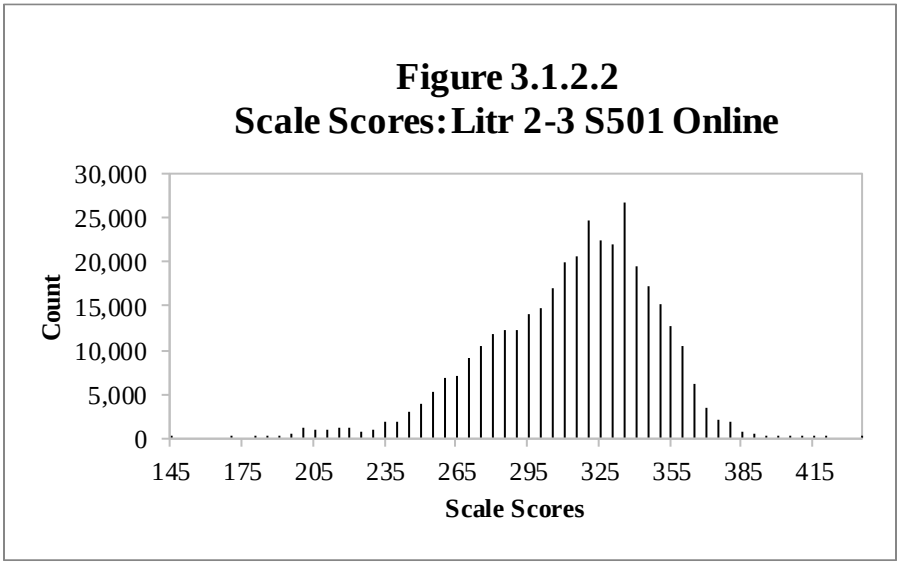


3.1.2.2 *Grades 2–3*

Table 3.1.2.2

Scale Score Descriptive Statistics: Litr 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	184,069	146	412	306.30	32.53
3	182,392	146	437	323.70	34.02
Total	366,461	146	437	314.96	34.40

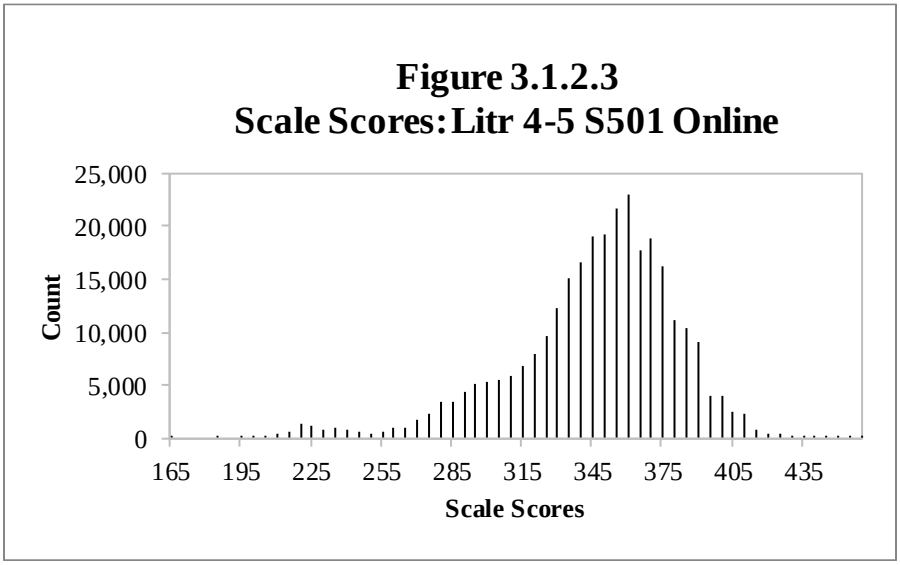


3.1.2.3 *Grades 4–5*

Table 3.1.2.3

Scale Score Descriptive Statistics: Litr 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	164,493	165	461	344.59	35.22
5	133,355	188	460	350.28	37.18
Total	297,848	165	461	347.14	36.22

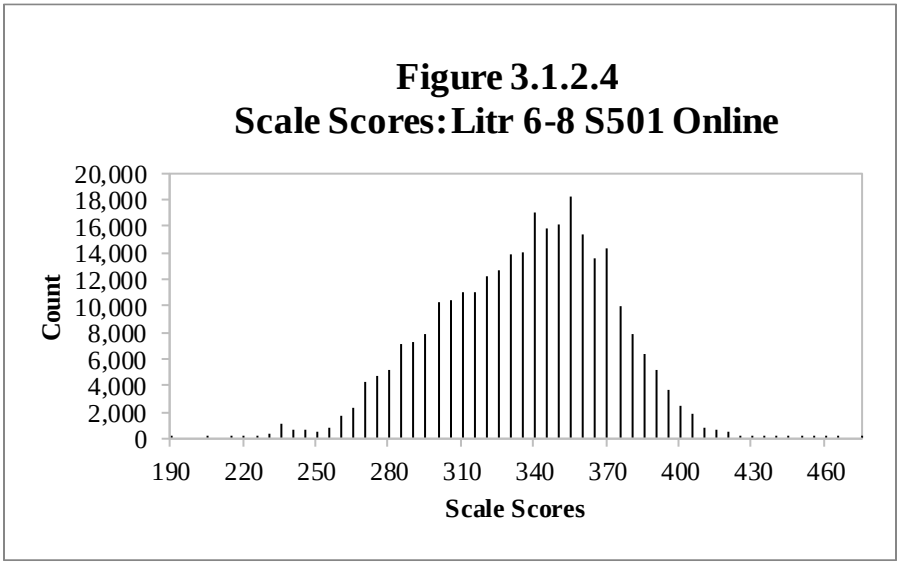


3.1.2.4 *Grades 6–8*

Table 3.1.2.4

Scale Score Descriptive Statistics: Litr 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	108,666	194	442	332.75	32.17
7	97,506	194	463	337.60	34.96
8	84,774	194	475	342.50	37.82
Total	290,946	194	475	337.21	35.05

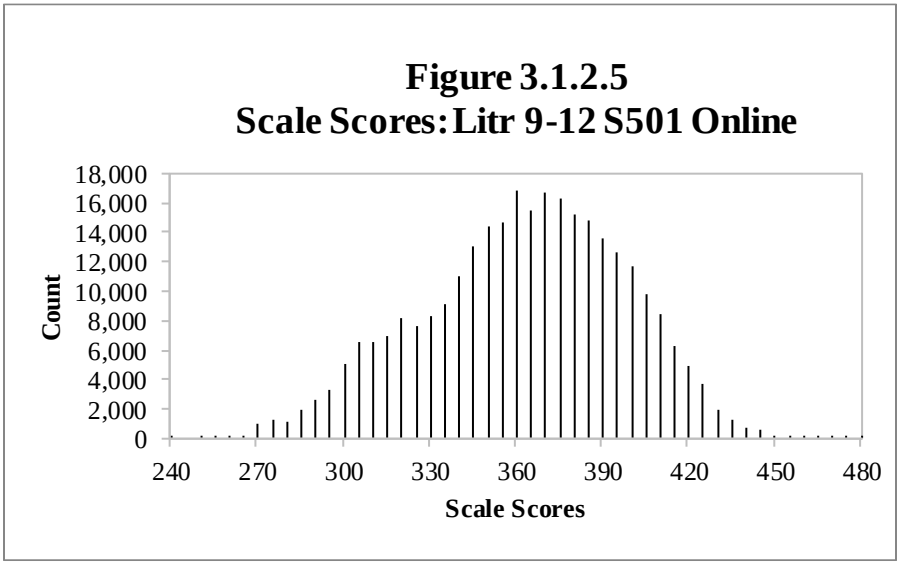


3.1.2.5 *Grades 9–12*

Table 3.1.2.5

Scale Score Descriptive Statistics: Litr 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	97,938	243	471	354.48	36.65
10	77,134	254	480	364.79	34.34
11	64,104	243	480	371.41	33.05
12	55,634	252	480	373.06	32.37
Total	294,810	243	480	364.36	35.32

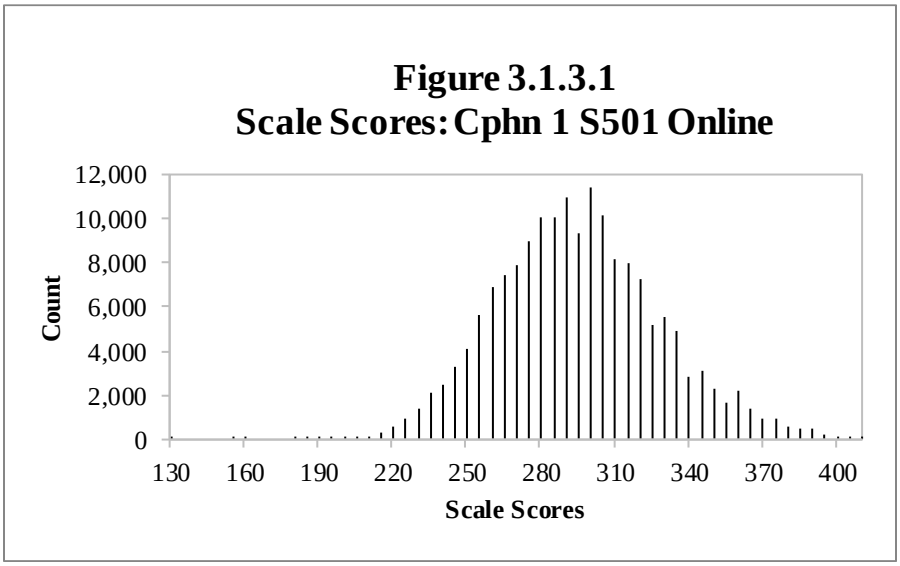


3.1.3 **Comprehension**

3.1.3.1 *Grade 1*

Table 3.1.3.1
Scale Score Descriptive Statistics: Cphn 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	170,589	130	411	297.08	33.13
Total	170,589	130	411	297.08	33.13

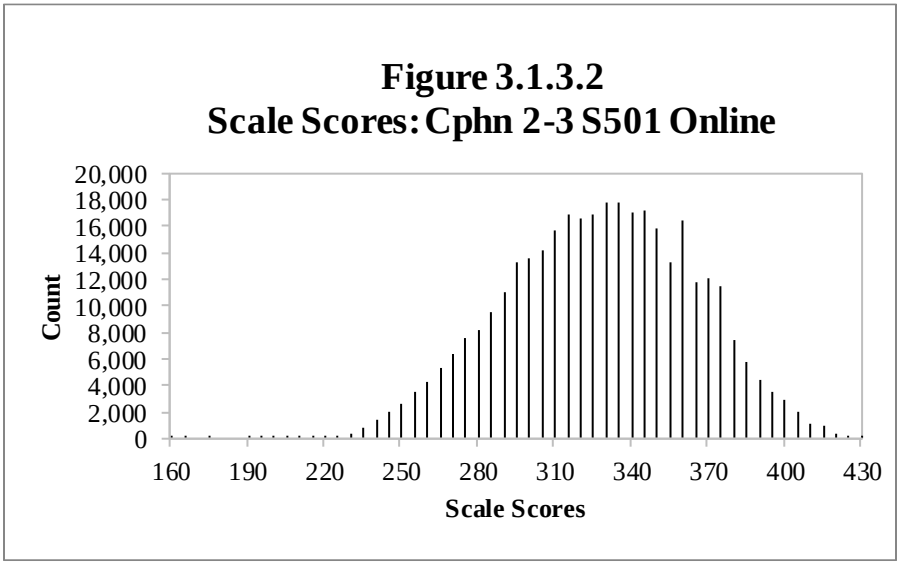


3.1.3.2 *Grades 2–3*

Table 3.1.3.2

Scale Score Descriptive Statistics: Cphn 2-3 S501 Online

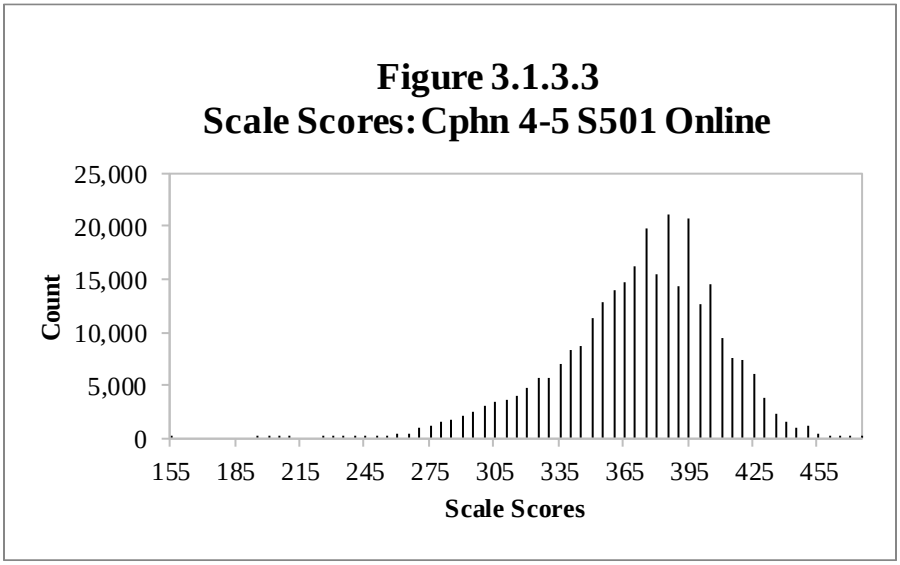
Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	175,334	164	430	320.69	33.66
3	174,482	160	430	338.36	37.67
Total	349,816	160	430	329.50	36.80



3.1.3.3 *Grades 4–5*

Table 3.1.3.3
Scale Score Descriptive Statistics: Cphn 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	163,186	197	478	371.62	34.71
5	131,842	159	478	376.33	37.63
Total	295,028	159	478	373.73	36.12



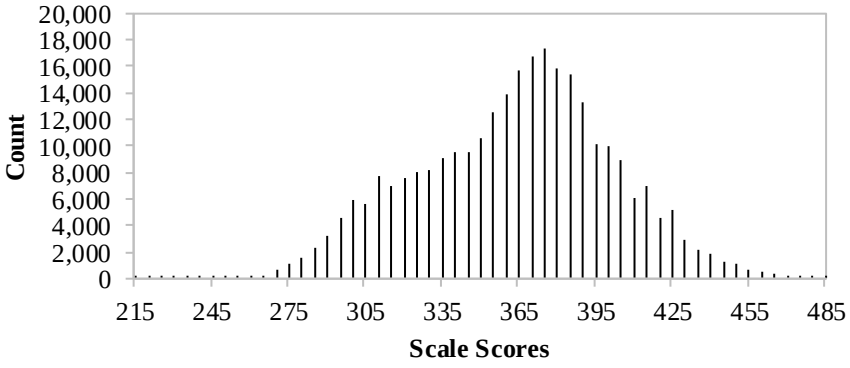
3.1.3.4 *Grades 6–8*

Table 3.1.3.4

Scale Score Descriptive Statistics: Cphn 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	106,592	222	488	360.76	34.22
7	95,881	223	488	365.66	38.16
8	83,806	216	488	371.47	41.79
Total	286,279	216	488	365.54	38.13

Figure 3.1.3.4
Scale Scores: Cphn 6-8 S501 Online



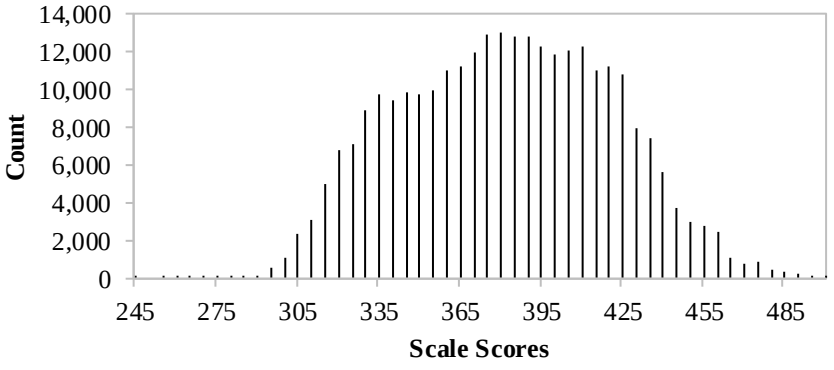
3.1.3.5 *Grades 9–12*

Table 3.1.3.5

Scale Score Descriptive Statistics: Cphn 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	95,312	263	502	374.24	38.79
10	75,461	267	502	384.21	38.96
11	62,705	267	502	390.84	38.25
12	54,599	247	502	392.52	37.54
Total	288,077	247	502	383.93	39.20

Figure 3.1.3.5
Scale Scores: Cphn 9-12 S501 Online

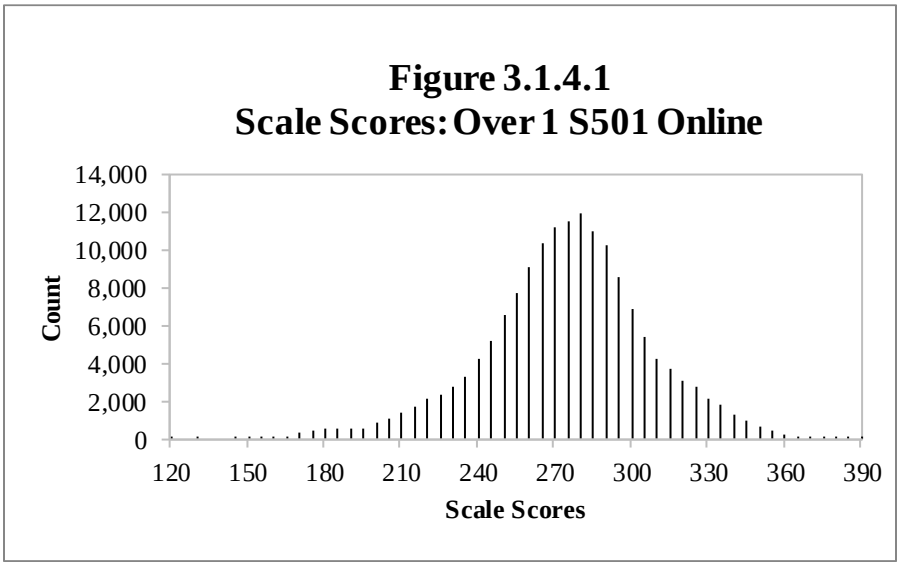


3.1.4 **Overall**

3.1.4.1 *Grade 1*

Table 3.1.4.1
Scale Score Descriptive Statistics: Over 1 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
1	160,535	120	392	276.11	32.34
Total	160,535	120	392	276.11	32.34

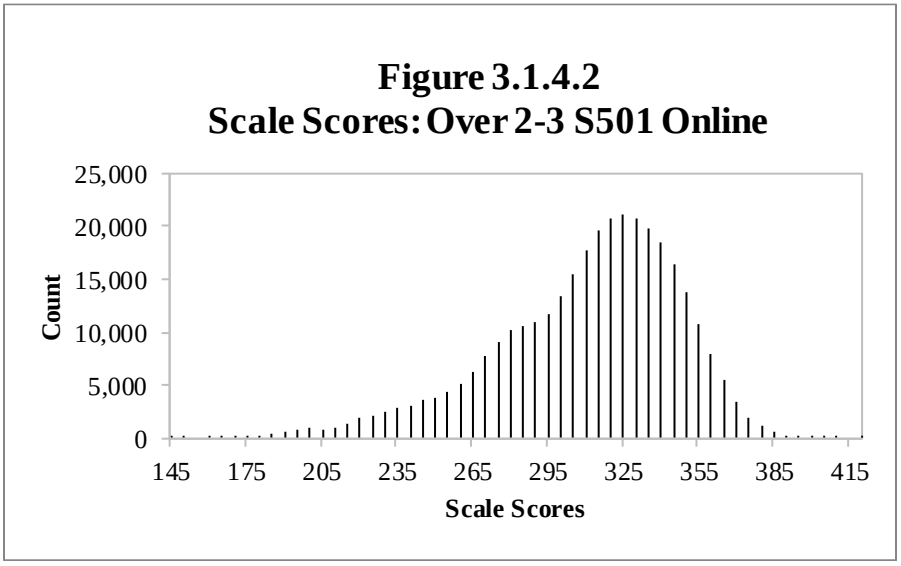


3.1.4.2 *Grades 2–3*

Table 3.1.4.2

Scale Score Descriptive Statistics: Over 2-3 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
2	165,651	153	402	303.27	34.68
3	165,992	146	423	321.45	36.51
Total	331,643	146	423	312.37	36.75

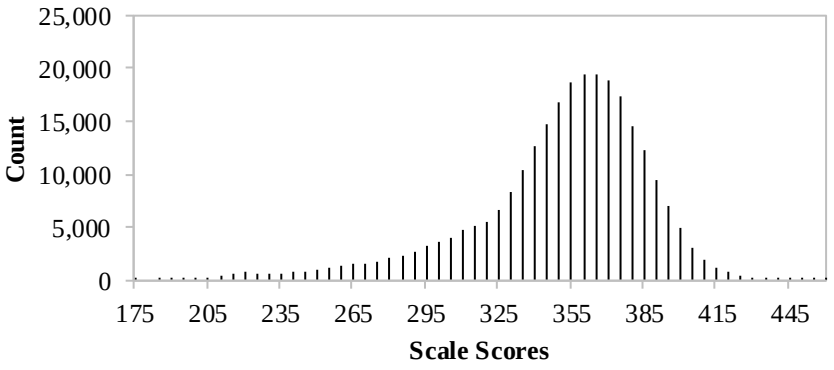


3.1.4.3 *Grades 4–5*

Table 3.1.4.3
Scale Score Descriptive Statistics: Over 4-5 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
4	147,625	175	449	349.87	35.73
5	120,064	189	462	354.89	38.25
Total	267,689	175	462	352.12	36.97

Figure 3.1.4.3
Scale Scores: Over 4-5 S501 Online



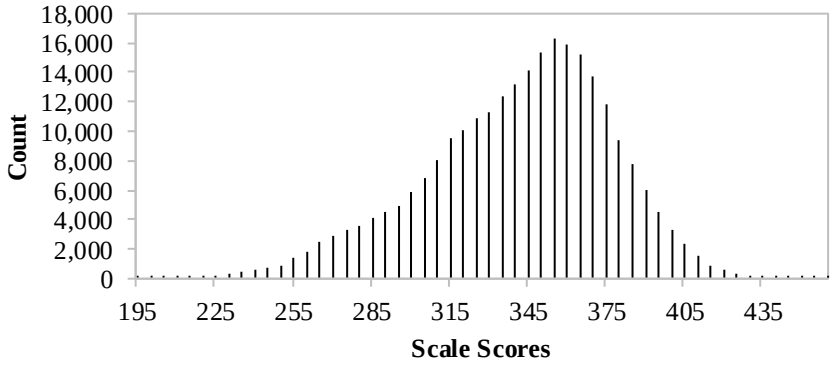
3.1.4.4 *Grades 6–8*

Table 3.1.4.4

Scale Score Descriptive Statistics: Over 6-8 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
6	96,583	204	447	339.56	32.16
7	86,658	205	461	343.49	35.81
8	75,984	196	463	347.90	39.10
Total	259,225	196	463	343.32	35.69

Figure 3.1.4.4
Scale Scores: Over 6-8 S501 Online

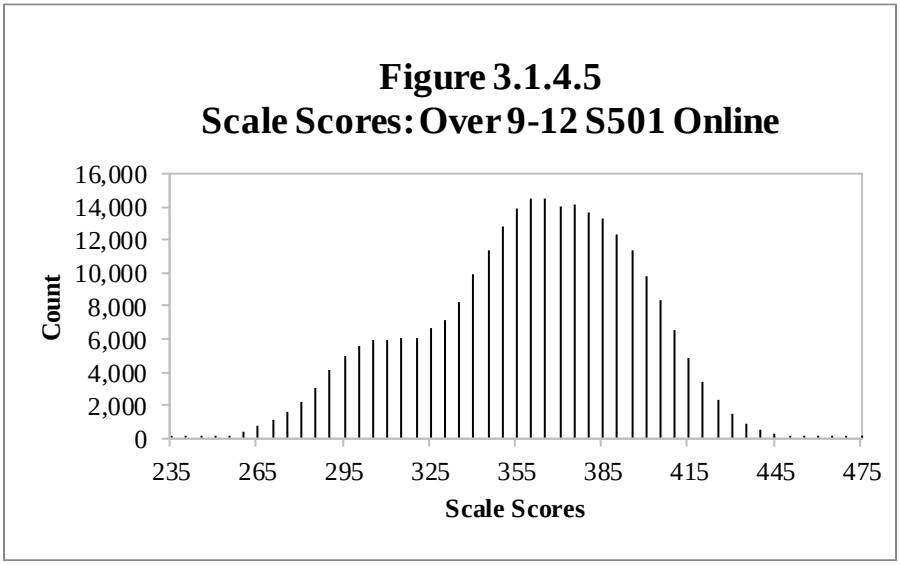


3.1.4.5 *Grades 9–12*

Table 3.1.4.5

Scale Score Descriptive Statistics: Over 9-12 S501 Online

Grade	No. of Students	Min.	Max.	Mean	Std. Dev.
9	87,417	239	465	351.18	38.08
10	69,142	249	471	361.40	35.92
11	57,295	246	474	367.96	34.59
12	50,306	252	476	369.22	33.81
Total	264,160	239	476	360.93	36.74



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

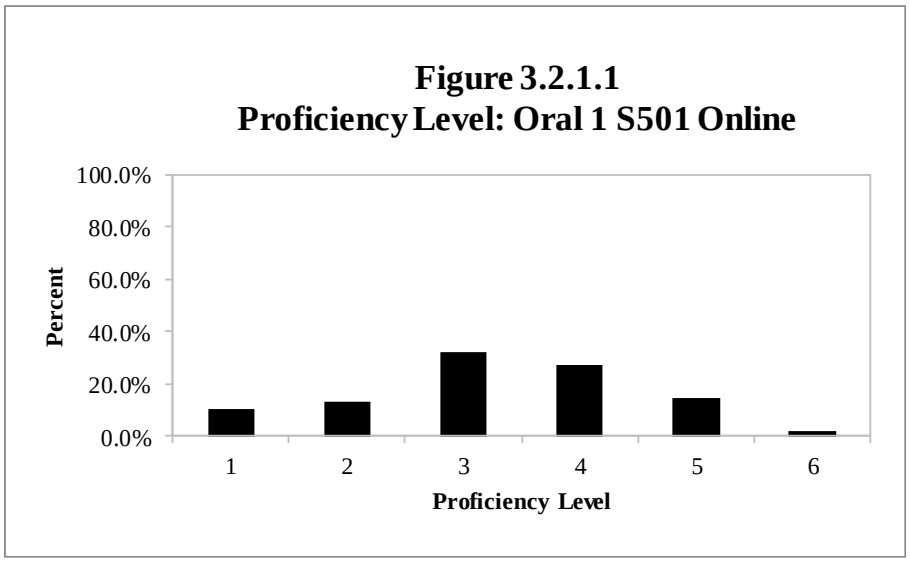
3.2.1 **Oral**

3.2.1.1 *Grade 1*

Table 3.2.1.1

Proficiency Level Distribution: Oral 1 S501 Online

	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	17,652	10.64%	17,652	10.64%
2	21,522	12.97%	21,522	12.97%
3	53,619	32.31%	53,619	32.31%
4	45,374	27.34%	45,374	27.34%
5	24,283	14.63%	24,283	14.63%
6	3,485	2.10%	3,485	2.10%
Total	165,935	100.00%	165,935	100.00%



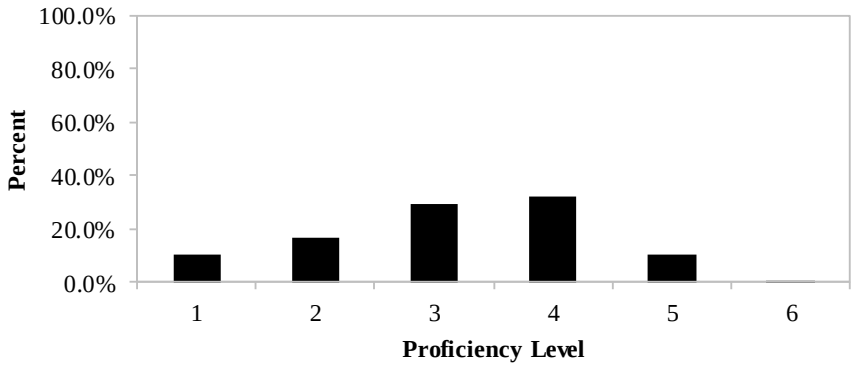
3.2.1.2 *Grades 2–3*

Table 3.2.1.2

Proficiency Level Distribution: Oral 2-3 S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	18,767	10.82%	16,435	9.47%	35,202	10.15%
2	32,934	19.00%	25,916	14.94%	58,850	16.97%
3	54,968	31.70%	48,114	27.74%	103,082	29.72%
4	48,483	27.96%	63,020	36.33%	111,503	32.15%
5	17,043	9.83%	18,960	10.93%	36,003	10.38%
6	1,186	0.68%	1,028	0.59%	2,214	0.64%
Total	173,381	100.00%	173,473	100.00%	346,854	100.00%

Figure 3.2.1.2
Proficiency Level: Oral 2-3 S501 Online



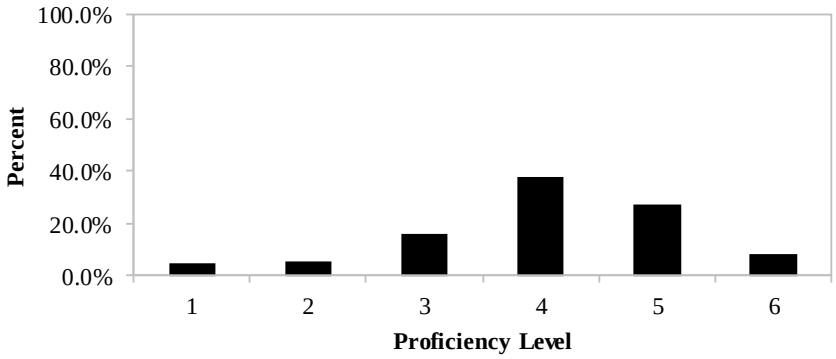
3.2.1.3 *Grades 4–5*

Table 3.2.1.3

Proficiency Level Distribution: Oral 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	6,230	3.82%	7,232	5.48%	13,462	4.56%
2	8,326	5.10%	7,476	5.66%	15,802	5.35%
3	25,671	15.74%	22,706	17.20%	48,377	16.39%
4	60,868	37.32%	50,092	37.94%	110,960	37.60%
5	46,332	28.41%	35,065	26.56%	81,397	27.58%
6	15,675	9.61%	9,464	7.17%	25,139	8.52%
Total	163,102	100.00%	132,035	100.00%	295,137	100.00%

Figure 3.2.1.3
Proficiency Level: Oral 4-5 S501 Online

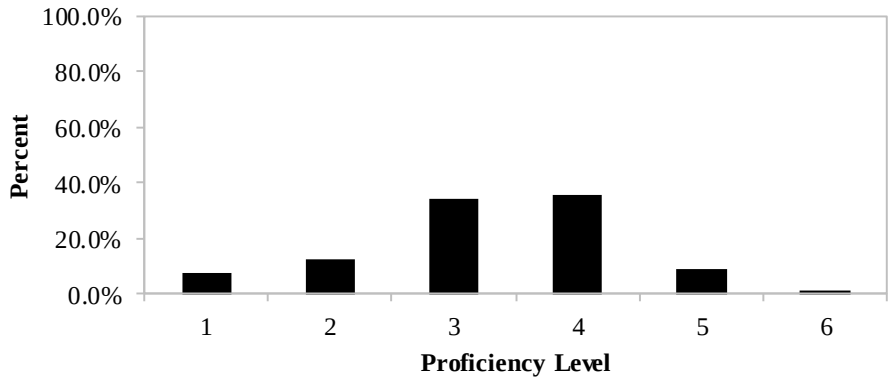


3.2.1.4 *Grades 6–8*

Table 3.2.1.4
Proficiency Level Distribution: Oral 6-8 S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	5,767	5.42%	7,973	8.33%	8,834	10.47%	22,574	7.88%
2	11,448	10.75%	13,041	13.63%	12,365	14.66%	36,854	12.86%
3	37,171	34.91%	33,298	34.80%	27,014	32.02%	97,483	34.02%
4	41,256	38.74%	31,860	33.30%	28,257	33.49%	101,373	35.38%
5	9,778	9.18%	8,397	8.78%	6,815	8.08%	24,990	8.72%
6	1,070	1.00%	1,120	1.17%	1,088	1.29%	3,278	1.14%
Total	106,490	100.00%	95,689	100.00%	84,373	100.00%	286,552	100.00%

Figure 3.2.1.4
Proficiency Level: Oral 6-8 S501 Online



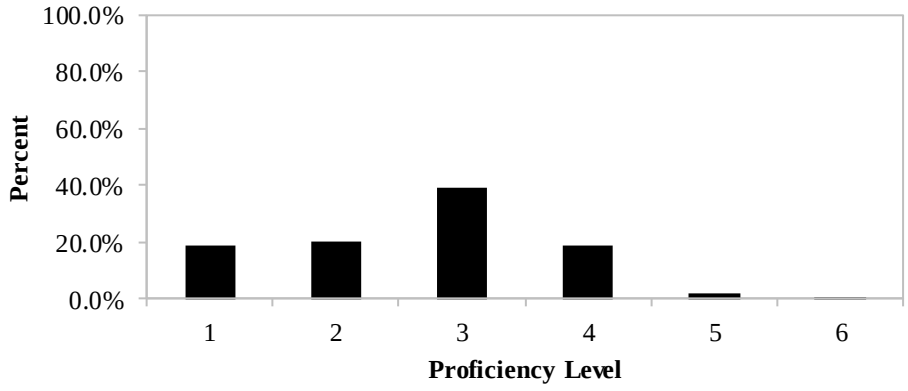
3.2.1.5 *Grades 9–12*

Table 3.2.1.5

Proficiency Level Distribution: Oral 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	20,106	20.94%	13,763	18.08%	10,810	17.13%	10,094	18.36%	54,773	18.87%
2	21,286	22.16%	14,993	19.70%	12,124	19.21%	10,594	19.27%	58,997	20.32%
3	33,411	34.79%	30,207	39.68%	26,651	42.22%	23,966	43.58%	114,235	39.35%
4	18,761	19.53%	15,150	19.90%	11,903	18.86%	9,308	16.93%	55,122	18.99%
5	2,231	2.32%	1,796	2.36%	1,459	2.31%	920	1.67%	6,406	2.21%
6	245	0.26%	210	0.28%	176	0.28%	108	0.20%	739	0.25%
Total	96,040	100.00%	76,119	100.00%	63,123	100.00%	54,990	100.00%	290,272	100.00%

Figure 3.2.1.5
Proficiency Level: Oral 9-12 S501 Online



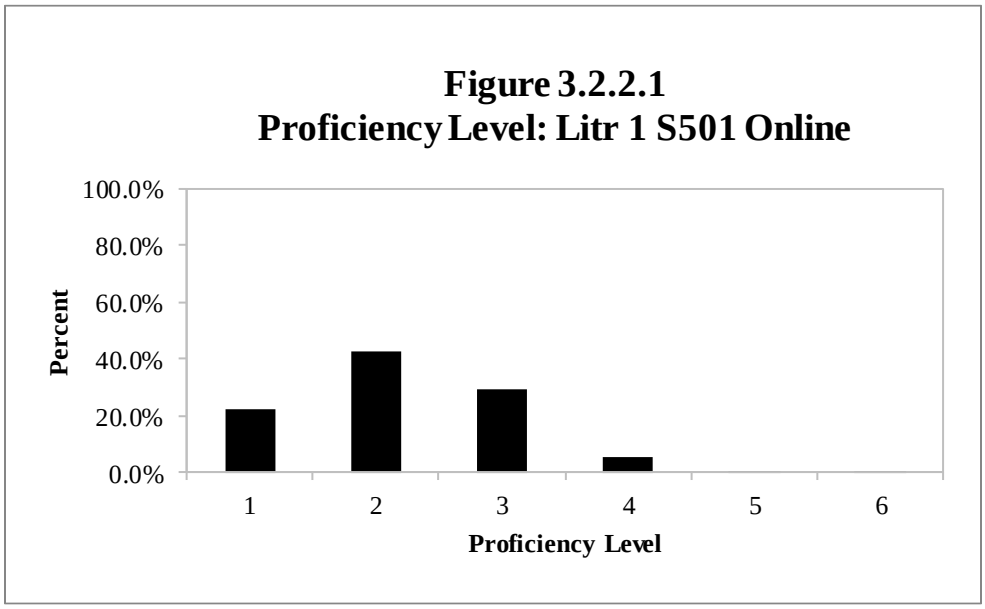
3.2.2 **Literacy**

3.2.2.1 *Grade 1*

Table 3.2.2.1

Proficiency Level Distribution: Litr 1 S501 Online

	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	39,835	22.17%	39,835	22.17%
2	76,171	42.39%	76,171	42.39%
3	52,189	29.04%	52,189	29.04%
4	9,786	5.45%	9,786	5.45%
5	1,584	0.88%	1,584	0.88%
6	132	0.07%	132	0.07%
Total	179,697	100.00%	179,697	100.00%



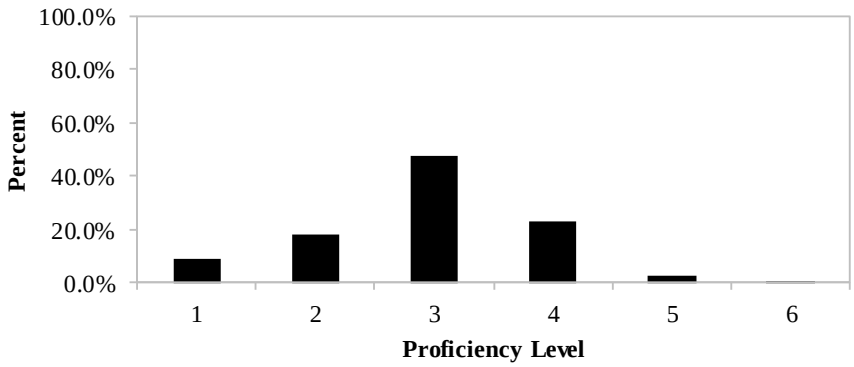
3.2.2.2 *Grades 2–3*

Table 3.2.2.2

Proficiency Level Distribution: Litr 2-3 S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	17,551	9.54%	15,265	8.37%	32,816	8.95%
2	39,394	21.40%	27,032	14.82%	66,426	18.13%
3	89,854	48.82%	83,958	46.03%	173,812	47.43%
4	34,170	18.56%	50,003	27.42%	84,173	22.97%
5	2,910	1.58%	5,884	3.23%	8,794	2.40%
6	190	0.10%	250	0.14%	440	0.12%
Total	184,069	100.00%	182,392	100.00%	366,461	100.00%

Figure 3.2.2.2
Proficiency Level: Litr 2-3 S501 Online

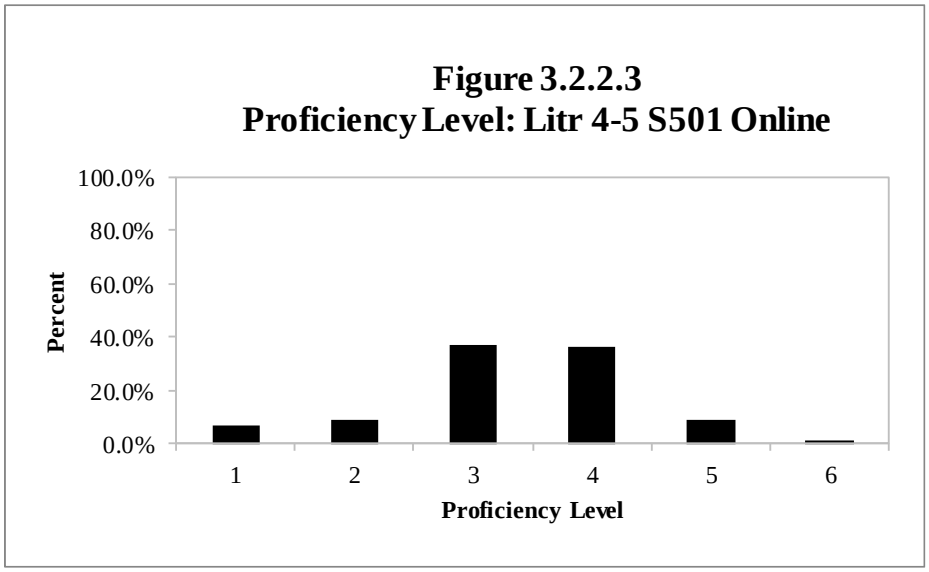


3.2.2.3 *Grades 4–5*

Table 3.2.2.3

Proficiency Level Distribution: Litr 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	Percent	Count	Percent	Count	Percent
1	10,786	6.56%	10,231	7.67%	21,017	7.06%
2	13,427	8.16%	12,774	9.58%	26,201	8.80%
3	63,013	38.31%	48,452	36.33%	111,465	37.42%
4	60,551	36.81%	47,622	35.71%	108,173	36.32%
5	13,972	8.49%	12,317	9.24%	26,289	8.83%
6	2,744	1.67%	1,959	1.47%	4,703	1.58%
Total	164,493	100.00%	133,355	100.00%	297,848	100.00%



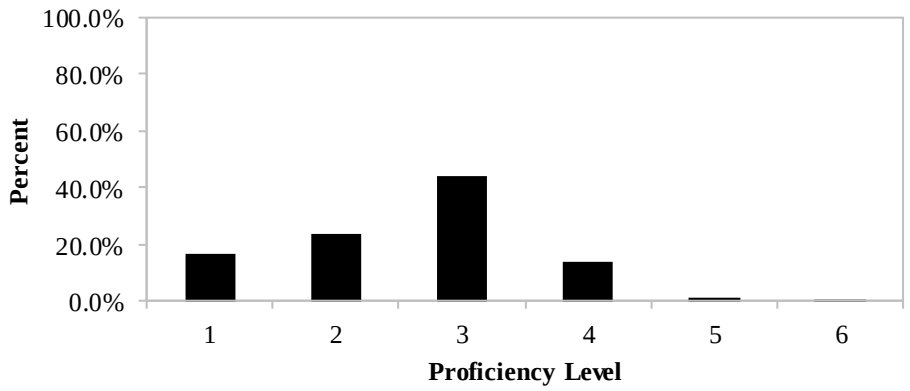
3.2.2.4 *Grades 6–8*

Table 3.2.2.4

Proficiency Level Distribution: Litr 6-8 S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	15,411	14.18%	15,882	16.29%	16,794	19.81%	48,087	16.53%
2	24,911	22.92%	24,335	24.96%	20,027	23.62%	69,273	23.81%
3	53,711	49.43%	42,562	43.65%	32,436	38.26%	128,709	44.24%
4	13,790	12.69%	13,379	13.72%	14,115	16.65%	41,284	14.19%
5	779	0.72%	1,294	1.33%	1,321	1.56%	3,394	1.17%
6	64	0.06%	54	0.06%	81	0.10%	199	0.07%
Total	108,666	100.00%	97,506	100.00%	84,774	100.00%	290,946	100.00%

Figure 3.2.2.4
Proficiency Level: Litr 6-8 S501 Online

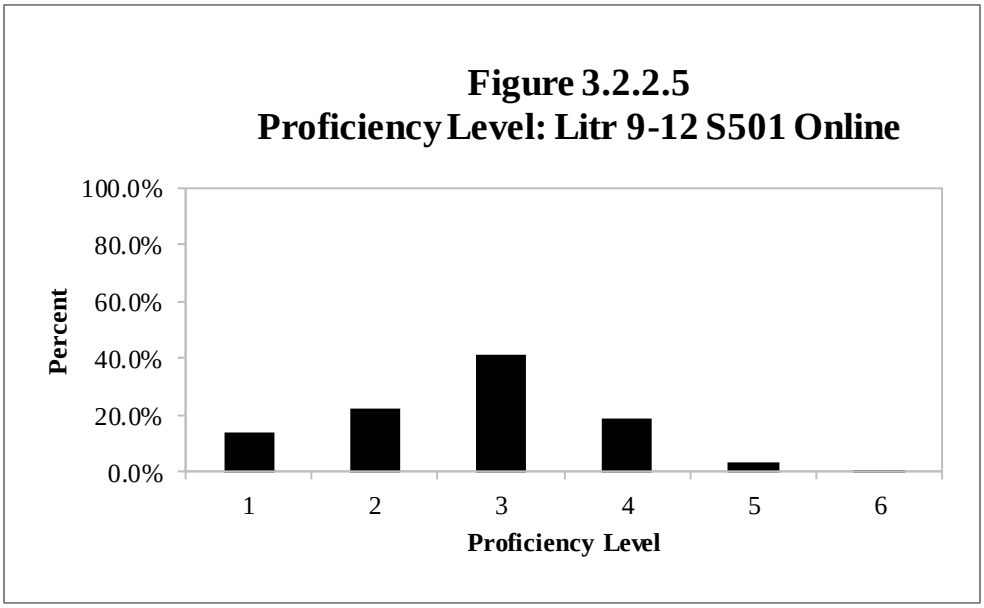


3.2.2.5 *Grades 9–12*

Table 3.2.2.5

Proficiency Level Distribution: Litr 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	16,961	17.32%	9,579	12.42%	7,024	10.96%	6,798	12.22%	40,362	13.69%
2	20,708	21.14%	16,399	21.26%	14,392	22.45%	14,481	26.03%	65,980	22.38%
3	38,439	39.25%	31,941	41.41%	27,290	42.57%	23,902	42.96%	121,572	41.24%
4	18,024	18.40%	16,098	20.87%	13,124	20.47%	9,214	16.56%	56,460	19.15%
5	3,629	3.71%	3,021	3.92%	2,243	3.50%	1,227	2.21%	10,120	3.43%
6	177	0.18%	96	0.12%	31	0.05%	12	0.02%	316	0.11%
Total	97,938	100.00%	77,134	100.00%	64,104	100.00%	55,634	100.00%	294,810	100.00%



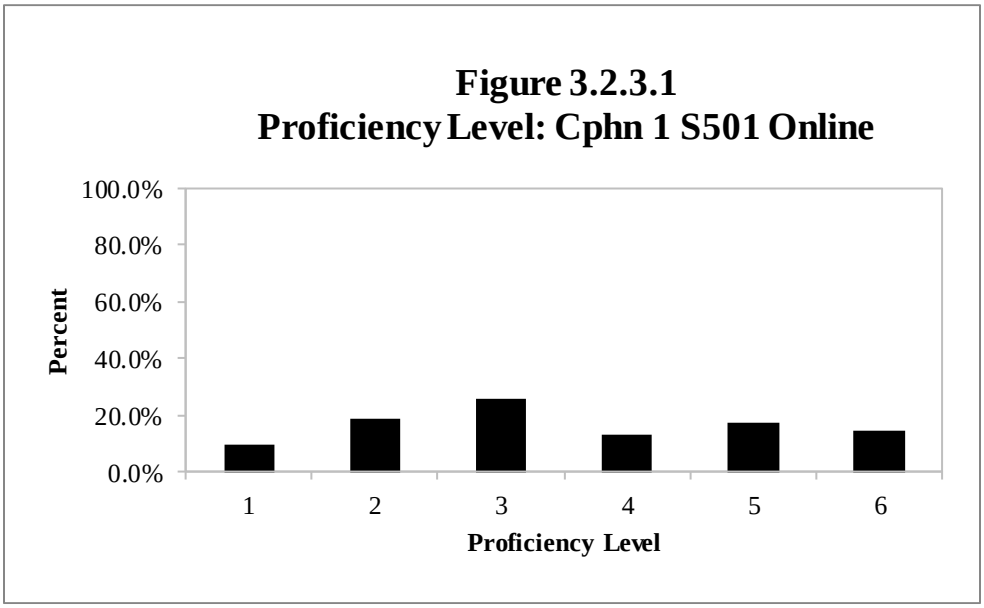
3.2.3 **Comprehension**

3.2.3.1 *Grade 1*

Table 3.2.3.1

Proficiency Level Distribution: Cphn 1 S501 Online

	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	16,876	9.89%	16,876	9.89%
2	31,849	18.67%	31,849	18.67%
3	44,059	25.83%	44,059	25.83%
4	23,043	13.51%	23,043	13.51%
5	29,925	17.54%	29,925	17.54%
6	24,837	14.56%	24,837	14.56%
Total	170,589	100.00%	170,589	100.00%



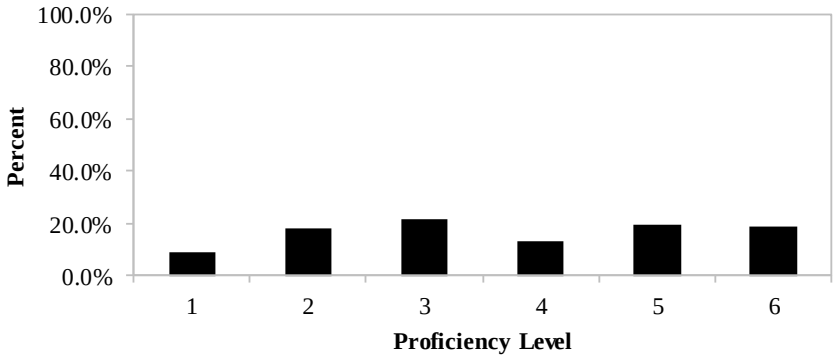
3.2.3.2 *Grades 2–3*

Table 3.2.3.2

Proficiency Level Distribution: Cphn 2-3 S501 Online

Level	Grade 2		Grade 3		Total	
	Count	Percent	Count	Percent	Count	Percent
1	13,645	7.78%	17,847	10.23%	31,492	9.00%
2	34,561	19.71%	29,115	16.69%	63,676	18.20%
3	39,888	22.75%	34,972	20.04%	74,860	21.40%
4	24,952	14.23%	20,322	11.65%	45,274	12.94%
5	33,433	19.07%	35,872	20.56%	69,305	19.81%
6	28,855	16.46%	36,354	20.84%	65,209	18.64%
Total	175,334	100.00%	174,482	100.00%	349,816	100.00%

Figure 3.2.3.2
Proficiency Level: Cphn 2-3 S501 Online



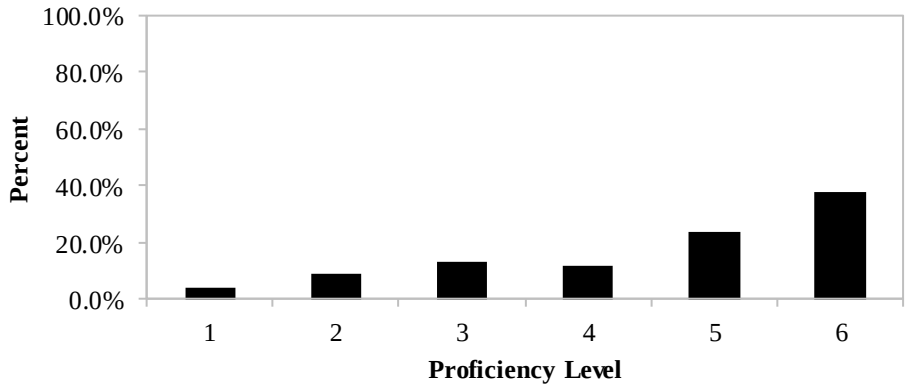
3.2.3.3 *Grades 4–5*

Table 3.2.3.3

Proficiency Level Distribution: Cphn 4-5 S501 Online

	Grade 4		Grade 5		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	4,983	3.05%	7,889	5.98%	12,872	4.36%
2	13,615	8.34%	12,098	9.18%	25,713	8.72%
3	21,759	13.33%	17,922	13.59%	39,681	13.45%
4	18,710	11.47%	16,786	12.73%	35,496	12.03%
5	40,032	24.53%	30,340	23.01%	70,372	23.85%
6	64,087	39.27%	46,807	35.50%	110,894	37.59%
Total	163,186	100.00%	131,842	100.00%	295,028	100.00%

Figure 3.2.3.3
Proficiency Level: Cphn 4-5 S501 Online



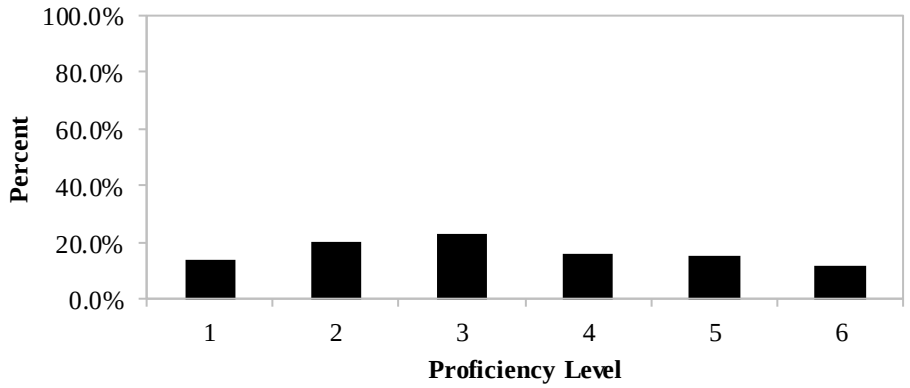
3.2.3.4 *Grades 6–8*

Table 3.2.3.4

Proficiency Level Distribution: Cphn 6-8 S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	11,701	10.98%	13,800	14.39%	14,432	17.22%	39,933	13.95%
2	21,703	20.36%	19,484	20.32%	16,182	19.31%	57,369	20.04%
3	26,028	24.42%	22,508	23.47%	17,139	20.45%	65,675	22.94%
4	18,726	17.57%	15,056	15.70%	11,816	14.10%	45,598	15.93%
5	18,164	17.04%	13,590	14.17%	12,040	14.37%	43,794	15.30%
6	10,270	9.63%	11,443	11.93%	12,197	14.55%	33,910	11.85%
Total	106,592	100.00%	95,881	100.00%	83,806	100.00%	286,279	100.00%

Figure 3.2.3.4
Proficiency Level: Cphn 6-8 S501 Online

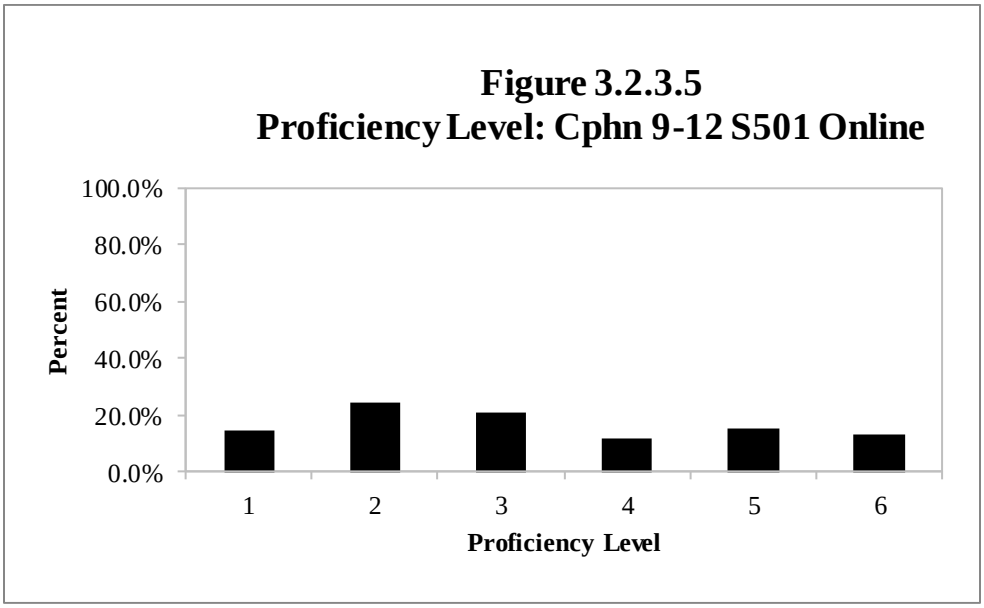


3.2.3.5 *Grades 9–12*

Table 3.2.3.5

Proficiency Level Distribution: Cphn 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	15,373	16.13%	10,617	14.07%	8,206	13.09%	7,633	13.98%	41,829	14.52%
2	25,333	26.58%	17,652	23.39%	14,328	22.85%	13,055	23.91%	70,368	24.43%
3	19,647	20.61%	15,998	21.20%	13,114	20.91%	11,575	21.20%	60,334	20.94%
4	10,358	10.87%	8,979	11.90%	7,371	11.76%	7,152	13.10%	33,860	11.75%
5	12,956	13.59%	11,453	15.18%	10,700	17.06%	8,905	16.31%	44,014	15.28%
6	11,645	12.22%	10,762	14.26%	8,986	14.33%	6,279	11.50%	37,672	13.08%
Total	95,312	100.00%	75,461	100.00%	62,705	100.00%	54,599	100.00%	288,077	100.00%



3.2.4 **Overall**

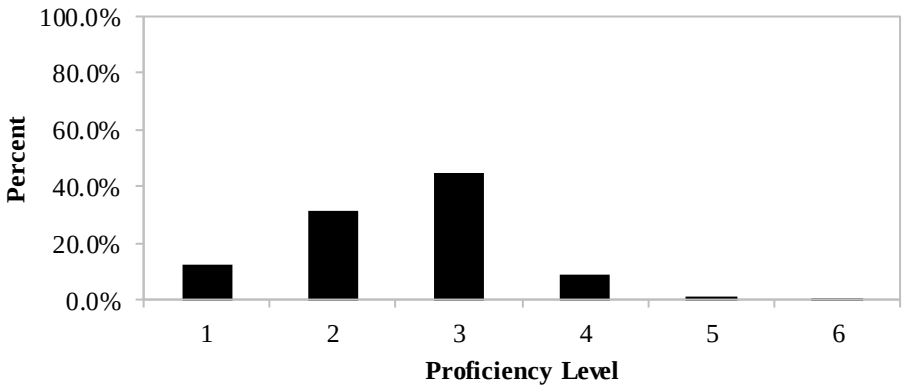
3.2.4.1 *Grade 1*

Table 3.2.4.1

Proficiency Level Distribution: Over 1 S501 Online

	Grade 1		Total	
Level	**Count**	**Percent**	**Count**	**Percent**
1	20,548	12.80%	20,548	12.80%
2	50,453	31.43%	50,453	31.43%
3	72,193	44.97%	72,193	44.97%
4	14,561	9.07%	14,561	9.07%
5	2,635	1.64%	2,635	1.64%
6	145	0.09%	145	0.09%
Total	160,535	100.00%	160,535	100.00%

Figure 3.2.4.1
Proficiency Level: Over 1 S501 Online



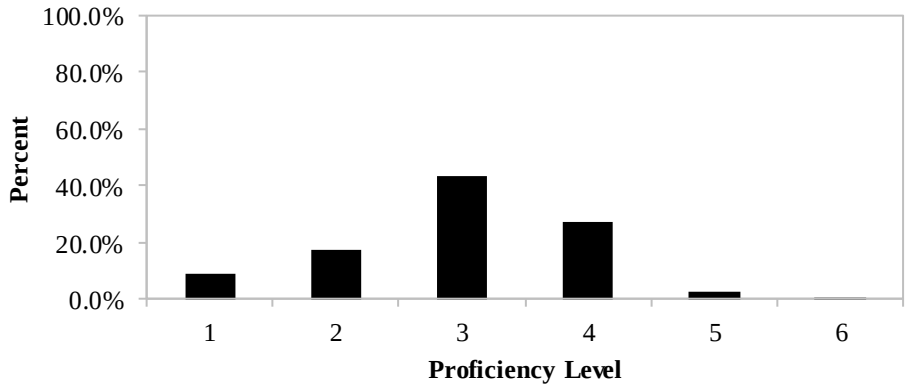
3.2.4.2 *Grades 2–3*

Table 3.2.4.2

Proficiency Level Distribution: Over 2-3 S501 Online

	Grade 2		Grade 3		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	15,215	9.18%	14,110	8.50%	29,325	8.84%
2	34,694	20.94%	23,068	13.90%	57,762	17.42%
3	75,027	45.29%	69,967	42.15%	144,994	43.72%
4	37,016	22.35%	52,997	31.93%	90,013	27.14%
5	3,599	2.17%	5,756	3.47%	9,355	2.82%
6	100	0.06%	94	0.06%	194	0.06%
Total	165,651	100.00%	165,992	100.00%	331,643	100.00%

Figure 3.2.4.2
Proficiency Level: Over 2-3 S501 Online



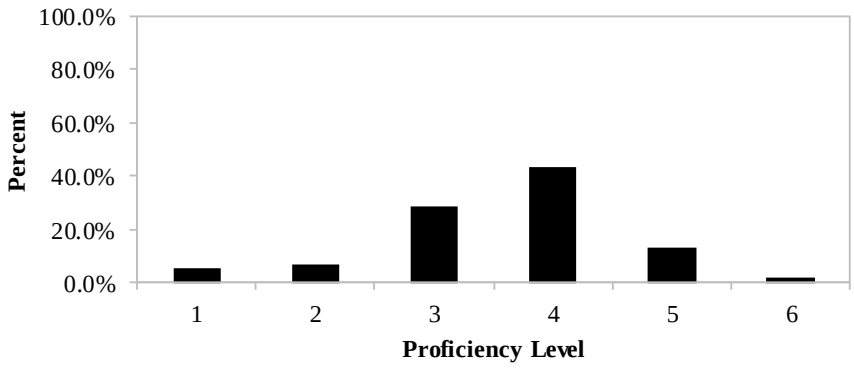
3.2.4.3 *Grades 4–5*

Table 3.2.4.3

Proficiency Level Distribution: Over 4-5 S501 Online

Level	Grade 4		Grade 5		Total	
	Count	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	7,483	5.07%	7,715	6.43%	15,198	5.68%
2	9,771	6.62%	9,581	7.98%	19,352	7.23%
3	43,058	29.17%	34,440	28.68%	77,498	28.95%
4	64,888	43.95%	51,445	42.85%	116,333	43.46%
5	19,522	13.22%	15,074	12.55%	34,596	12.92%
6	2,903	1.97%	1,809	1.51%	4,712	1.76%
Total	147,625	100.00%	120,064	100.00%	267,689	100.00%

Figure 3.2.4.3
Proficiency Level: Over 4-5 S501 Online



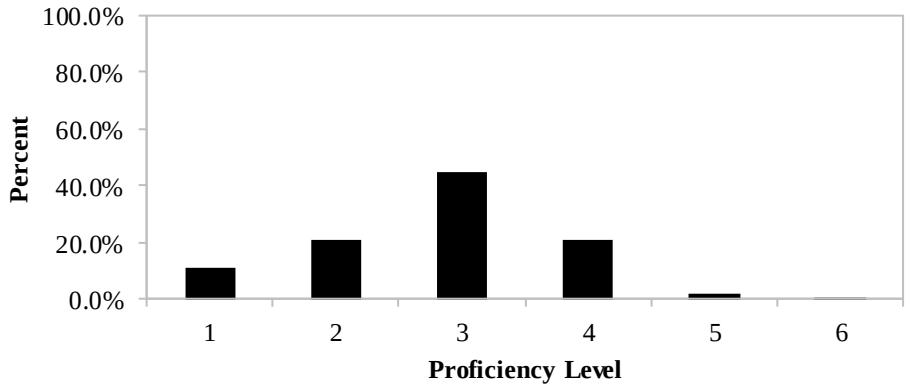
3.2.4.4 *Grades 6–8*

Table 3.2.4.4

Proficiency Level Distribution: Over 6-8 S501 Online

	Grade 6		Grade 7		Grade 8		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	8,233	8.52%	10,163	11.73%	11,063	14.56%	29,459	11.36%
2	18,758	19.42%	18,420	21.26%	16,362	21.53%	53,540	20.65%
3	48,387	50.10%	38,374	44.28%	29,891	39.34%	116,652	45.00%
4	19,963	20.67%	18,122	20.91%	16,801	22.11%	54,886	21.17%
5	1,166	1.21%	1,514	1.75%	1,778	2.34%	4,458	1.72%
6	76	0.08%	65	0.08%	89	0.12%	230	0.09%
Total	96,583	100.00%	86,658	100.00%	75,984	100.00%	259,225	100.00%

Figure 3.2.4.4
Proficiency Level: Over 6-8 S501 Online

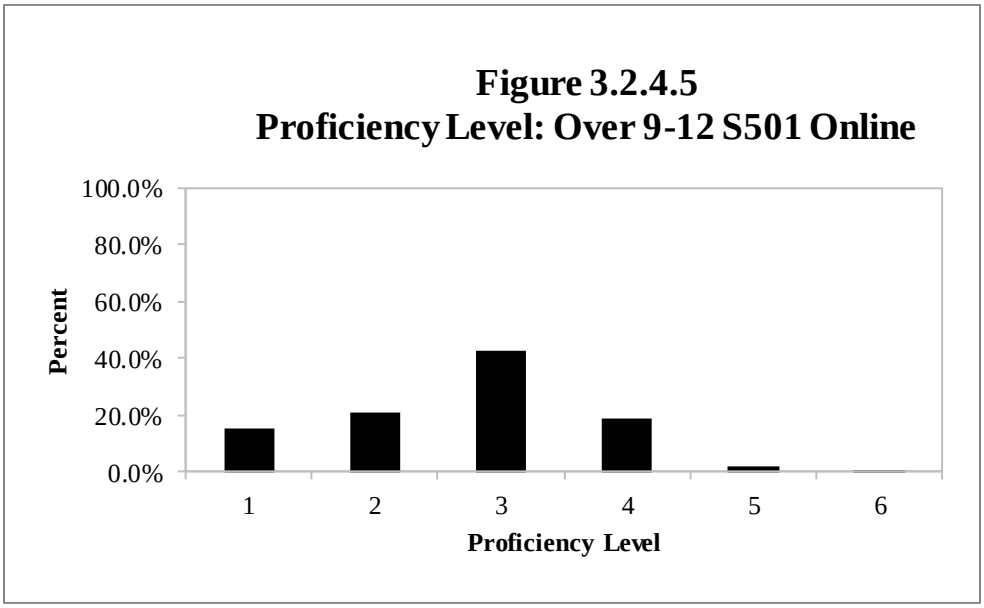


3.2.4.5 *Grades 9–12*

Table 3.2.4.5

Proficiency Level Distribution: Over 9-12 S501 Online

	Grade 9		Grade 10		Grade 11		Grade 12		Total	
Level	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
1	16,153	18.48%	9,660	13.97%	7,209	12.58%	6,932	13.78%	39,954	15.12%
2	17,605	20.14%	13,962	20.19%	11,847	20.68%	12,293	24.44%	55,707	21.09%
3	35,542	40.66%	29,923	43.28%	25,773	44.98%	22,346	44.42%	113,584	43.00%
4	16,073	18.39%	14,023	20.28%	11,245	19.63%	8,099	16.10%	49,440	18.72%
5	1,954	2.24%	1,526	2.21%	1,199	2.09%	631	1.25%	5,310	2.01%
6	90	0.10%	48	0.07%	22	0.04%	5	0.01%	165	0.06%
Total	87,417	100.00%	69,142	100.00%	57,295	100.00%	50,306	100.00%	264,160	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.

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.

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.

Section 4.2.2, Study of Technology-Enhanced Items, describes how these innovative types of items fit/enhance the construct of the Listening and Reading domains compared with typically used item types.

Section 4.2.3, Study of Differential Item Functioning by Disability Status, addresses whether items of Listening and Reading domains can be analyzed in a DIF procedure using disability status and whether items show bias against disability status or disability type.

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, English Learner Reclassification Study—Phase 1, addresses the validity of using the ACCESS test to reclassify EL learners for exiting from the supporting programs.

4.2. Annual Validity Studies

4.2.1. English Learner Reclassification Study—Phase 1

Kim, A., Ho, P., Chapman, M., & Cook, H. G. (2020a). *Examination of reclassification decisions made for K–12 English learners: Survey report of Delaware* (WIDA Internal Report). Madison, WI: WIDA at the Wisconsin Center for Education Research.

Kim, A., Ho, P., Chapman, M., & Cook, H. G. (2020b). *Examination of reclassification decisions made for K–12 English learners: Survey report of Pennsylvania* (WIDA Internal Report). Madison, WI: WIDA at the Wisconsin Center for Education Research.

A survey was conducted to investigate how ELs are reclassified across districts in select WIDA Consortium member states. Despite the high-stakes nature of the reclassification decision, little is known regarding the decision-making process across WIDA states. A pilot survey was distributed across districts in Vermont in spring of 2019; findings were used to update the main survey. The revised survey consisted of five sections: (1) educator background information, (2)

reclassification criteria, (3) reclassification procedures and decision-makers, (4) reclassification monitoring, and (5) perceived effectiveness of reclassification.

Two states—Delaware and Pennsylvania—were recruited for the main study (Kim, Ho, Chapman, & Cook, 2020a, 2020b). According to its reclassification policy, Delaware uses only English language proficiency assessment scores, whereas Pennsylvania uses both English language proficiency assessment scores and teacher judgments on students' classroom language proficiency. Online surveys were distributed across districts in September to October 2019. Collected data were primarily analyzed using descriptive analyses. Open-ended responses were qualitatively analyzed for emerging patterns.

Results from Pennsylvania indicated that EL reclassification criteria varied across districts (Kim et al., 2020b). The state's policy requires a minimum of two criteria for making reclassification decisions: ELs' scores on an English language proficiency assessment (ACCESS for ELLs) and educator input (standardized language use inventory). Findings indicated that over half of the districts (65%) used three or more criteria for EL reclassification, for example, students' writing samples, performance in content areas, and grade-point average. Such variability in the number and types of criteria could potentially result in ELs qualifying for reclassification in one district but not in others.

Survey findings also indicated that reclassification decisions were either made by a single decision-maker (37%) or through a reclassification meeting (46%) attended by several educators. In either case, district EL/Title III coordinators and EL/Bilingual program directors were often the primary decision-makers for EL reclassification. Although few educators believed that ELs were inappropriately reclassified, students' disability status was considered the main factor leading to inappropriate reclassification. Overall, these results suggest that the majority of Pennsylvania districts and schools exercise local autonomy regarding EL reclassification, creating wide variability in decision-making across districts. Furthermore, these findings from Phase 1 will guide Phase 2 of the study (see Phase 2 under Ongoing Research).

In the survey, educators shared their suggestions for improving EL reclassification. They requested more targeted training from the state. Examples included more training for content teachers, who were not as familiar with English language proficiency terminology and concepts, and more professional development on reclassifying ELs with disabilities. Some educators also believed that ACCESS for ELLs could be enhanced by ensuring that its Speaking domain better reflects students' actual speaking language ability.

4.2.2. Technology-Enhanced Items Study

Kim, A., Tywoniw, R. L., & Chapman, M. (2020). *Performance of Technology-Enhanced Items in Grades 1-12 English Language Proficiency Assessments*.

Technology-enhanced items (TEIs) are innovative computer-delivered items that require interactions with the test environment beyond traditional multiple-choice items (MCIs). This interactive nature allows TEIs to measure test constructs better than MCIs (Sireci & Zenisky, 2006). Examples of TEIs include hotspot and drag-and-drop items, which require test-takers to either click an area with text or images or drag the answer to designated zones. Despite the popularity of TEIs in computer-based assessments, there is little research that compares students' performance on TEIs vs. MCIs in English language proficiency (ELP) assessments. In addition, there is little understanding of how TEI innovations enhance accessibility of items for multilingual learners. Previous research on TEIs is limited to math and science domains (Crabtree, 2016), and research on TEIs in ELP contexts is rare, especially in K–12 settings.

This study examined ELs' performance on hotspot and drag-and-drop TEIs vs. MCIs on the Reading domain of ACCESS for ELLs. We analyzed 1.2 million ELs' scores on this domain across five grade-cluster levels: Grades 1, 2–3, 4–5, 6–8, and 9–12. The reading test measures students' academic reading development, a critical skill for academic success. The test included 24 to 30 MCIs per grade level, as well as several field test items, which were content-matched TEIs and MCIs. That is, these pairs shared the same content but differed in their response mode. Content-matched TEIs and MCIs were evaluated for standard item performance metrics such as difficulty, discrimination, and information using item response theory modeling. In addition, item efficiency was measured using the amount of item information provided in relation to item duration. Moreover, to examine how TEIs affect the accessibility of the test, we examined ELs' use of several online accessibility features: color options (overlay and contrast), a highlighter tool, a line guide tool, a magnifier tool, and a help button for general and tool help.

Overall, TEIs were found to be slightly more difficult than content-matched MCIs, but they did not differ in discriminative power. The information provided by TEIs to the overall test varied by grade level, typically being more informative for ELs in higher grade levels or proficiency levels. Regarding item efficiency, TEIs took more time for learners to respond to and generally had longer item duration. Yet, TEIs were on average more efficient than MCIs in Grades 6–8, providing more information for these select grades. Furthermore, TEIs elicited more use of accessibility features across all test-takers, especially of the highlighter and line-guide tools.

These quantitative results were augmented with qualitative analysis of reading item design features to further understand the intersection of technology enhancement and measurement of K–12 ELs' reading proficiency. Taken together, these findings provide insights for further development of TEIs in online ELP assessments for multilingual learners that embrace the interactivity of TEIs and mitigate potential difficulties. A report on this study is forthcoming.

4.2.3. Study of Differential Item Functioning by Disability Status

Bishop, K., Walker, C., Gocer Sahin, S., and Akanda, M. (2020). *DIF study by disability status in EL assessment*. WIDA Technical Report.

This study examines differential item functioning (DIF) of ACCESS Online items by disability status. The purpose of this study is to investigate whether items are disadvantaging disability groups and to provide information about the appropriateness and fairness of ACCESS Online test items.

. To assess fairness as a lack of bias, item performance is tested via differential item functioning (DIF). This study examined how disability status relates to different ability distribution and disability groups in ACCESS's online multistage Listening and Reading tests.

The findings showed inconsistency of variance among disability groups, which leads to a heterogeneity issue in performing the DIF procedure with a disability group vs. nondisability group. Within the disability group, those on the autism spectrum showed the highest variance with the lowest mean scores. DIF-flagged items were fewer and were also balanced between the disability and nondisability group in Listening and Reading tests. Since the general grouping of all disability groups together did not meet the homogeneity assumption of the DIF procedure, this study suggests performing DIF in each disability group separately against a nondisability group. Another next step is to examine proper accommodation tool use.

5 Reliability

In accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014), in interpreting test scores, it is important to evaluate their reliability, as the interpretation of test scores depends on assumptions that students exhibit some degree of consistency in their scores across independent administrations of the same testing procedure. It is expected 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 particular sample of items and tasks used to assess students. Furthermore, because it is assumed that all items on such a test measure some aspect of the domain of interest, it is expected 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 questions or tasks vary from one sample of test questions 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 as well as how the test scores will be interpreted. There are several ways to collect reliability data and to estimate reliability, many 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 compliant with critical element 4.1 of the ESSA 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). Reliability of domain score is presented first, followed by reliability of composite scores.

ACCESS Listening, Reading, Writing, and Speaking scores are used to determine the English language proficiency of students based on students' test scores in each of the four domains. Therefore, the main concern in interpreting the ACCESS test scores is how consistent the scores of the students would be over replications of the same testing procedure in each domain. We use **internal-consistency reliability statistics** to address this question (Section 5.1). Additionally, for the Writing and Speaking domains, inconsistency in test scores may be introduced by

different raters as a potential source of variation. The **interrater agreement** in scoring Writing and Speaking tasks is reported in Section 5.2, to examine how consistent the scores of the students would be if their responses were scored by different raters. Since an item response theory–based method is used in estimating students' latent scores, we also examine the amount of **measurement error** in students' scores using conditional standard error of measurement (Section 5.3). Lastly, in Section 5.4, we evaluate the reliability of classification into WIDA proficiency levels (the most important interpretation of the test scores) in terms of the **accuracy and consistency** of the classification decisions made based on the students' domain test scores. Detailed descriptions of the methods, data sources, and procedures are presented in each subsection.

ACCESS composite scores are used to describe the English language proficiency of students in the respective composites. Therefore, the most important concern in interpreting the ACCESS composite scores is how consistent the composites scores of these students would be over replications of the same testing procedure. We use internal consistency reliability statistics to address this question, and results are provided in Section 5.5. In addition, we examine conditional standard error of measurement of the composites in Section 5.6. Lastly, we evaluate the reliability of classification in terms of the accuracy and the consistency of the decisions made about students' level of English language proficiency based on their composite scores in Section 5.7. Detailed descriptions of the methods, data sources, and procedures are presented in each section.

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 (parallel test form reliability). Given that the abilities being measured are assumed to be constant for each student over two administrations of alternate forms, the more variation found across the two administrations, the more evidence for lower reliability. In this case, the sources of inconsistency across the two administrations taken together are called "measurement error." Measurement error is considered to be random and to occur by chance. For example, there may be some kinds of knowledge and skills assessed by some items or tasks that affect students' scores, but which are not part of what the test intends to measure.

Unless students take two alternate versions of the same test, test reliability cannot be calculated directly. Thus, it is usually estimated from student responses to a single form of the test. Methods used to estimate reliability using test scores from a single test administration are modeled from classical test theory and are referred to as estimates of ***internal consistency***. Internal consistency reliability statistics are a good estimate of alternate-forms reliability statistics, providing an estimate of the consistency of the performance of students across items within a test. The most common index of internal consistency reliability is referred to as Cronbach's alpha (Cronbach, 1951), which is a lower bound estimate of test reliability. Conceptually, it may be thought of as

the correlation obtained between performances on two halves of the test, if every possible way of dividing the test items in two were attempted. Because Cronbach's alpha is a correlation of all possible pairs of test items, Cronbach's alpha may be low if some items are measuring something other than what most of the other items are measuring (and thus leading to inconsistent student performances). In this way, Cronbach's alpha expresses how well the items and tasks on a test appear to measure the same ability. The Cronbach's alpha coefficient of internal consistency ranges from 0 to 1. If scores are assigned to students by a completely random process (i.e., scores are not correlated or share no covariance), then the reliability estimate is very close to 0. If scores assigned to students are perfectly consistent (i.e., scores have high covariances), then the internal consistency coefficient will approach 1.

Reliability statistics such as the Cronbach alpha coefficient of internal consistency are affected by the number of test items or test score points that may be awarded. That is, all things being equal, the greater the number of items measuring similar abilities 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, they are affected by the distribution of abilities measured by the test within the specific group of students tested. If the students in the group are nearly equal in the abilities measured by the test (i.e., are very homogeneous in 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 examinees tested than the test itself. On the other hand, if the students in the group differ widely in the abilities the test measures (i.e., are very heterogeneous in the ability distribution), 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, it is widely recognized that reliability can be as much a function of the test items and tasks as of the sample of students tested. That is, the exact same test can produce widely disparate reliability indices based on the distribution of the group of students. Therefore, when interpreting estimates of internal consistency, it is wise to keep in mind the specific set of test items and the distribution of ability in the group of students used in the estimation.

Interrater Agreement

A potential source of variance in students' scores on the productive domains of ACCESS (Writing and Speaking) lies in the behavior of raters. ACCESS scoring procedures and steps taken to provide rater training and consistency are described elsewhere in this report (see Part 1, Section 3.2.2). The **interrater agreement** rates in scoring Writing and Speaking tasks are reported in Section 5.2. These values examine how consistent the scores of the students would be if their responses were scored by different groups of raters. Detailed descriptions of the methods, data sources, and procedures are presented in the section.

Measurement Error

In addition to evaluating test reliability in terms of estimates of internal consistency, the amount of measurement error in students' test scores is commonly addressed in two different ways in educational and psychological testing. One way is to hypothesize that there is an error-free measure of students' true ability, skills, or proficiency. In classical test theory, it is referred to as the true score. True score is a theoretical value, so it is not a known quantity. Rather, it is viewed as the hypothetical average score over repeated replications of the same testing condition. Under the assumption of classical test theory, the error of measurement over replication of a testing condition provides an estimate of the amount of variability we would expect from students' true scores. In practical testing contexts, it is generally not possible to replicate a testing condition (i.e., have students take the same test form over and over again), so it is not possible to estimate the standard error of the students' scores using a repeated measure design. Instead, the average error of measurement over the population of students who take the test is estimated and used as an indication of the amount of variation we would expect in any individual student's score. This statistic is referred to as the *standard error of measurement* (SEM). It 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, a confidence interval can be constructed 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 standard error of measurement at specific scores using item response theory (IRT). IRT addresses reliability using the information function, which indicates the precision with which student performances on items and tasks can be used to estimate the latent ability of each student. The square root of the inverse of the information function at any point on the latent ability distribution is the conditional standard error of measurement (CSEM). The CSEM provides information about the amount of error we would expect in any student's score at that point on the underlying latent ability scale and is expressed in terms of the latent score metric (i.e., the IRT metric for expressing student ability, as opposed to the raw score). In addition, using IRT, indices analogous to traditional reliability coefficients such as Cronbach's alpha can be estimated from the test information 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 level 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 the WIDA proficiency levels (American Educational Research Association et al., 2014). The questions we want to

answer are different from the questions answered by the reliability coefficient. Instead of looking at the reliability of a specific student score, we want to know how consistently the classifications are being made about students when placed by their test results into a smaller number of proficiency levels. One way to approach this question is to estimate the degree to which classification decisions we are making on the basis of the students' observed test scores agree with the classification decisions we would make based on students' theoretical true score. This estimate is known as decision accuracy. A second way to approach this question is to estimate the degree to which classification decisions we are making on the basis of the students' test scores agree with the classification decisions we would make based on students' scores on a different edition of the test. This estimate is known as decision consistency.

5.1 Reliability of Domain Scores

Listening and Reading

Internal consistency statistics based on classical test theory are applicable only on a fixed-test-length test where all students take the same set of test items (Thissen, 2000). For Listening and Reading domains that are computer adaptive, traditional internal consistency statistics cannot be applied because not all students take the same set of items. We estimate reliability for Listening and Reading by grade-level cluster using an IRT-based marginal reliability method derived by Thissen (2000). Unlike the traditional internal consistency statistics that are based on students' raw score, the marginal reliability method uses students' modeled latent scores and student distribution in its estimation. However, the marginal reliability can be interpreted like other traditional internal consistency statistics such as Cronbach's coefficient alpha (Thissen, 2000).

The formula for IRT marginal reliability method developed by Thissen (2000) is

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

where

$\bar{\rho}$ is the average reliability
σ_θ^2 is the variance of the distribution of the student ability measures
$CSEM_{observed}^2$ is the squared observed conditional standard errors of measurement for each student

The IRT marginal reliability can be derived directly (Thissen, 2000); however, it is computationally intensive. Since this estimate is equivalent to the Rasch person reliability coefficient (Linacre, 1999) which is readily available in Winsteps, for purposes of efficiency WIDA chose to present the Rasch person reliability as the test reliability estimate for the Listening and Reading domains. The Rasch person reliability coefficient is an estimate of the ratio of "true measure variance" to "observed measure variance" (Linacre, 1999). To obtain these values, item parameters and population student data were used as inputs in the Winsteps program.

In the tables below that present reliability information for Listening and Reading, we provide the Rasch person reliability coefficient for ACCESS Online. For these two domains, the first table provides the Rasch person reliability coefficient (labeled as 'Rasch Reliability Coefficient' in the table) for all students. Each row in the table represents a grade-level cluster, and values for the numbers of students, numbers of items, and the reliability estimate are provided for each grade-level cluster. The second table for each domain provides the same information for the population of female students and the population of male students. The third table provides information by ethnicity, for Hispanic and non-Hispanic test-takers, and the fourth table provides information for the population of students who have an individualized education plan (IEP).

The Listening Rasch person reliability computed for all students ranged from 0.82 to 0.86 across the grade clusters. The Listening Rasch person reliability ranged from 0.83 to 0.87 for male students; 0.81 to 0.86 for female students; 0.82 to 0.87 for Hispanic students; 0.79 to 0.85 for non-Hispanic students; and 0.81 to 0.89 for students with an IEP.

The Reading Rasch person reliability computed for all students ranged from 0.88 to 0.91. The Reading Rasch person reliability ranged from 0.88 to 0.92 for male students; 0.88 to 0.91 for female students; 0.86 to 0.91 for Hispanic students; 0.88 to 0.92 for non-Hispanic students; and 0.83 to 0.88 for students with an IEP.

Writing and Speaking

Cronbach's coefficient alpha is widely used as an estimate of reliability, particularly of the internal consistency of test items, and this statistic is appropriate for the Writing and Speaking fixed forms. Conceptually, it may be thought of as the correlation obtained between performances on two halves of the test, if every possible way of dividing the test tasks in two were attempted. Thus, Cronbach's alpha may be low if some items are measuring something other than what the majority of the items are measuring. In this way, Cronbach's alpha expresses how well the items and tasks on a test appear to measure the same ability.

The formula for Cronbach's alpha is

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_t^2}\right]$$

where

n = number of items
σ_i^2 = variance of score on item i
σ_t^2 = variance of total score

For Writing and Speaking, tables in this section also present the standard error of measurement which provides a single value for estimating the errors of measurement in students' scores using classical test theory. It is a function of two statistics: the Cronbach's alpha of the test and the (observed) standard deviation (SD) of the test scores in the student population, and it is on the raw score metric. It is calculated as

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

Since the standard error of measurement is an estimate of the standard deviation of the distribution of measurement errors, SEM can be used to create a band around a student's observed 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 score has a 68% probability of falling within the band extending from the observed score minus 1 SEM to the observed score plus 1 SEM. Since SEMs are expressed on the raw score metric, it is wise to keep the range of the raw score distribution in mind when interpreting the SEM. Raw score statistics by domains are reported in Section 2.

In the tables below that present reliability information for Writing and Speaking, we provide the number of tasks, Cronbach's alpha, and SEM for all students and for subgroups as required by the ESSA Peer Review so that the reliability estimates of the subgroups can be compared with those computed based on all students. For these domains, the first table provides Cronbach's alpha and the SEM for all students. Each row in the table represents a specific grade cluster and test form. For each form, the numbers of students, numbers of tasks, Cronbach's alpha, and SEM are provided. The second table for each domain provides the same information for the population of female students and the population of male students. The third table provides information by ethnicity, for Hispanic and non-Hispanic test-takers, and the fourth table provides information for the population of students who have an IEP.

Note that students' Writing reported scores are based on student performances on only two tasks starting with Online Series 501, and Cronbach's alpha for the Writing domain may be lower than when estimated on the basis of three tasks, as in earlier series.

Writing Tier A: The Writing Tier A Cronbach's alpha computed for all students ranged from 0.80 to 0.88. The Writing Tier A Cronbach's alpha ranged from 0.81 to 0.88 for male students; 0.79 to 0.87 for female students; 0.80 to 0.88 for Hispanic students; 0.78 to 0.86 for non-Hispanic students; and 0.76 to 0.86 for students with an IEP.

Writing Tier B/C: The Writing Tier B/C Cronbach's alpha computed for all students ranged from 0.56 to 0.70. The Writing Tier B/C Cronbach's alpha ranged from 0.60 to 0.71 for male students; 0.51 to 0.67 for female students; 0.57 to 0.70 for Hispanic students; 0.54 to 0.68 for non-Hispanic students; and 0.62 to 0.77 for students with an IEP.

Speaking Tier Pre-A: The Speaking Tier Pre-A Cronbach's alpha computed for all students ranged from 0.84 to 0.86. Cronbach's alpha ranged from 0.83 to 0.87 for male students; 0.84 to 0.86 for female students; 0.83 to 0.86 for Hispanic students; 0.83 to 0.88 for non-Hispanic students; and 0.83 to 0.91 for students with an IEP.

Speaking Tier A: The Speaking Tier A Cronbach's alpha computed for all students ranged from 0.81 to 0.84. Cronbach's alpha ranged from 0.80 to 0.84 for male students; 0.81 to 0.83 for female students; 0.81 to 0.84 for Hispanic students; 0.77 to 0.82 for non-Hispanic students; and 0.75 to 0.85 for students with an IEP.

Speaking Tier B/C: The Speaking Tier B/C Cronbach's alpha computed for all students ranged from 0.80 to 0.86. Cronbach's alpha ranged from 0.79 to 0.86 for male students; 0.80 to 0.86 for female students; 0.80 to 0.87 for Hispanic students; 0.79 to 0.83 for non-Hispanic students; and 0.81 to 0.87 for students with an IEP.

5.1.1 Listening

Table 5.1.1.1

Reliability: List S501 Online

Cluster	No. of Students	No. of Items	Rasch Reliability Estimate
1	176,572	54	0.86
2-3	366,603	54	0.86
4-5	315,715	54	0.82
6-8	306,619	54	0.85
9-12	309,545	54	0.85

Table 5.1.1.2

Reliability: List S501 Online by Gender

		Female		Male	
Cluster	No. of Items	No. of Students	Rasch Reliability Estimate	No. of Students	Rasch Reliability Estimate
1	54	80,990	0.86	90,826	0.87
2-3	54	167,269	0.85	189,385	0.86
4-5	54	140,628	0.81	167,096	0.83
6-8	54	130,761	0.85	167,193	0.86
9-12	54	131,413	0.84	168,926	0.85

Table 5.1.1.3

Reliability: List S501 Online by Ethnicity

		Hispanic		Other	
Cluster	No. of Items	No. of Students	Rasch Reliability Estimate	No. of Students	Rasch Reliability Estimate
1	54	113,284	0.87	56,930	0.85
2-3	54	240,337	0.87	113,095	0.84
4-5	54	216,195	0.82	84,955	0.79
6-8	54	209,947	0.85	78,874	0.84
9-12	54	204,323	0.85	86,725	0.84

Table 5.1.1.4

Reliability: List S501 Online by IEP Status

Cluster	No. of Students	No. of Items	Rasch Reliability Estimate
1	14,402	54	0.89
2-3	34,611	54	0.88
4-5	39,213	54	0.81
6-8	48,836	54	0.82
9-12	40,305	54	0.81

5.1.2 **Reading**

Table 5.1.2.1

Reliability: Read S501 Online

Cluster	No. of Students	No. of Items	Rasch Reliability Estimate
1	179,739	72	0.88
2-3	366,612	72	0.88
4-5	309,547	72	0.89
6-8	304,091	72	0.91
9-12	304,775	72	0.91

Table 5.1.2.2

Reliability: Read S501 Online by Gender

Cluster	No. of Items	Female		Male	
		No. of Students	Rasch Reliability Estimate	No. of Students	Rasch Reliability Estimate
1	72	82,091	0.88	92,738	0.88
2-3	72	166,352	0.88	190,181	0.88
4-5	72	137,038	0.88	164,533	0.89
6-8	72	128,769	0.91	166,679	0.91
9-12	72	128,417	0.91	167,220	0.92

Table 5.1.2.3

Reliability: Read S501 Online by Ethnicity

Cluster	No. of Items	Hispanic		Other	
		No. of Students	Rasch Reliability Estimate	No. of Students	Rasch Reliability Estimate
1	72	115,625	0.86	57,611	0.90
2-3	72	240,447	0.87	112,872	0.88
4-5	72	212,111	0.88	82,915	0.89
6-8	72	208,756	0.91	77,464	0.92
9-12	72	202,074	0.91	84,283	0.91

Table 5.1.2.4

Reliability: Read S501 Online by IEP Status

Cluster	No. of Students	No. of Items	Rasch Reliability Estimate
1	14,763	72	0.83
2-3	34,735	72	0.85
4-5	38,606	72	0.87
6-8	48,907	72	0.88
9-12	40,213	72	0.88

5.1.3 **Writing**

Table 5.1.3.1

Reliability: Writ S501 Online

Cluster	Tier	No. of Students	No. of Tasks	Cronbach's Alpha	SEM
1	A	158,459	2	0.802	1.184
	B/C	28,391	2	0.596	1.236
2-3	A	95,649	2	0.875	1.075
	B/C	290,488	2	0.696	1.011
4-5	A	49,912	2	0.869	1.098
	B/C	268,413	2	0.681	1.150
6-8	A	110,111	2	0.833	1.039
	B/C	201,973	2	0.564	1.244
9-12	A	114,168	2	0.851	1.187
	B/C	203,773	2	0.639	1.218

Table 5.1.3.2

Reliability: Writ S501 Online by Gender

Cluster	Tier	No. of Tasks	Female			Male		
			No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	A	2	71,244	0.785	1.185	82,735	0.810	1.181
	B/C	2	14,287	0.573	1.220	13,524	0.606	1.251
2-3	A	2	39,509	0.872	1.079	53,092	0.876	1.072
	B/C	2	136,321	0.666	0.987	146,694	0.707	1.030
4-5	A	2	20,129	0.865	1.103	28,399	0.871	1.091
	B/C	2	121,094	0.648	1.134	140,552	0.696	1.163
6-8	A	2	42,868	0.826	1.051	64,056	0.835	1.031
	B/C	2	89,075	0.507	1.274	107,151	0.604	1.213
9-12	A	2	44,364	0.843	1.189	66,562	0.853	1.184
	B/C	2	89,618	0.620	1.201	107,743	0.646	1.234

Table 5.1.3.3

Reliability: Writ S501 Online by Ethnicity

Cluster	Tier	No. of Tasks	Hispanic			Other		
			No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	A	2	107,049	0.804	1.190	45,542	0.780	1.165
	B/C	2	12,786	0.603	1.234	14,723	0.576	1.235
2-3	A	2	68,994	0.876	1.080	22,444	0.861	1.061
	B/C	2	183,709	0.701	1.021	97,003	0.670	0.989
4-5	A	2	35,557	0.869	1.102	10,906	0.851	1.097
	B/C	2	182,504	0.682	1.141	74,379	0.677	1.167
6-8	A	2	79,292	0.836	1.039	23,267	0.795	1.027
	B/C	2	134,814	0.573	1.210	56,326	0.542	1.322
9-12	A	2	81,502	0.852	1.185	24,757	0.820	1.184
	B/C	2	128,572	0.642	1.204	63,887	0.631	1.246

Table 5.1.3.4

Reliability: Writ S501 Online by IEP Status

Cluster	Tier	No. of Students	No. of Tasks	Cronbach's Alpha	SEM
1	A	14,470	2	0.829	1.173
	B/C	886	2	0.686	1.314
2-3	A	16,885	2	0.859	1.100
	B/C	19,715	2	0.772	1.108
4-5	A	11,854	2	0.827	1.105
	B/C	27,902	2	0.720	1.201
6-8	A	24,753	2	0.762	1.007
	B/C	25,389	2	0.619	1.203
9-12	A	17,509	2	0.801	1.182
	B/C	23,967	2	0.635	1.205

5.1.4 **Speaking**

Table 5.1.4.1

Reliability: Spek S501 Online

Cluster	Tier	No. of Students	No. of Tasks	Cronbach's Alpha	SEM
1	Pre-A	7,109	3	0.838	0.820
	A	67,864	6	0.827	1.342
	B/C	99,910	6	0.827	1.587
2-3	Pre-A	17,104	3	0.842	0.690
	A	82,157	6	0.810	1.225
	B/C	264,823	6	0.797	1.479
4-5	Pre-A	6,370	3	0.844	0.803
	A	31,669	6	0.807	1.336
	B/C	272,752	6	0.815	1.495
6-8	Pre-A	9,533	3	0.857	0.684
	A	62,225	6	0.835	1.229
	B/C	235,821	6	0.822	1.463
9-12	Pre-A	19,889	3	0.863	0.639
	A	128,946	6	0.837	1.296
	B/C	160,717	6	0.859	1.412

Table 5.1.4.2

Reliability: Spek S501 Online by Gender

Cluster	Tier	No. of Tasks	Female			Male		
			No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	Pre-A	3	2,725	0.840	0.805	4,156	0.837	0.826
	A	6	29,290	0.829	1.331	36,652	0.823	1.348
	B/C	6	48,395	0.828	1.587	48,921	0.821	1.583
2-3	Pre-A	3	6,998	0.855	0.660	9,580	0.832	0.708
	A	6	34,488	0.808	1.229	45,071	0.811	1.220
	B/C	6	124,803	0.795	1.484	133,105	0.794	1.474
4-5	Pre-A	3	2,625	0.851	0.773	3,559	0.834	0.824
	A	6	13,029	0.808	1.341	17,787	0.804	1.335
	B/C	6	122,934	0.810	1.497	142,886	0.819	1.490
6-8	Pre-A	3	3,926	0.852	0.680	5,348	0.859	0.686
	A	6	24,334	0.833	1.243	36,168	0.836	1.220
	B/C	6	102,068	0.824	1.481	126,916	0.820	1.449
9-12	Pre-A	3	7,856	0.851	0.626	11,507	0.869	0.645
	A	6	51,812	0.826	1.312	73,359	0.844	1.286
	B/C	6	70,998	0.855	1.433	84,647	0.861	1.394

Table 5.1.4.3

Reliability: Spek S501 Online by Ethnicity

Cluster	Tier	No. of Tasks	Hispanic			Other		
			No. of Students	Cronbach's Alpha	SEM	No. of Students	Cronbach's Alpha	SEM
1	Pre-A	3	5,111	0.834	0.835	1,644	0.850	0.755
	A	6	47,036	0.828	1.345	18,318	0.815	1.332
	B/C	6	60,131	0.828	1.579	36,344	0.819	1.597
2-3	Pre-A	3	12,508	0.840	0.703	3,724	0.838	0.616
	A	6	58,886	0.816	1.226	19,875	0.776	1.222
	B/C	6	167,026	0.801	1.473	88,820	0.787	1.492
4-5	Pre-A	3	4,644	0.845	0.798	1,031	0.832	0.732
	A	6	22,398	0.807	1.339	7,136	0.767	1.330
	B/C	6	185,678	0.816	1.486	75,342	0.808	1.513
6-8	Pre-A	3	7,178	0.851	0.686	1,349	0.866	0.589
	A	6	44,671	0.838	1.234	13,222	0.788	1.190
	B/C	6	158,831	0.825	1.449	64,296	0.807	1.496
9-12	Pre-A	3	14,684	0.857	0.653	3,572	0.882	0.516
	A	6	89,992	0.844	1.299	30,872	0.792	1.273
	B/C	6	99,992	0.866	1.408	51,751	0.832	1.419

Table 5.1.4.4

Reliability: Spek S501 Online by IEP Status

Cluster	Tier	No. of Students	No. of Tasks	Cronbach's Alpha	SEM
1	Pre-A	1,149	3	0.834	0.794
	A	8,085	6	0.824	1.341
	B/C	5,079	6	0.832	1.593
2-3	Pre-A	3,151	3	0.846	0.567
	A	14,249	6	0.783	1.209
	B/C	17,032	6	0.805	1.493
4-5	Pre-A	530	3	0.831	0.727
	A	7,612	6	0.748	1.307
	B/C	30,467	6	0.819	1.510
6-8	Pre-A	925	3	0.882	0.534
	A	13,959	6	0.812	1.167
	B/C	34,381	6	0.821	1.442
9-12	Pre-A	2,081	3	0.905	0.578
	A	21,318	6	0.852	1.258
	B/C	17,011	6	0.869	1.396

5.2 Interrater Agreement

For the Writing and Speaking tests, tables provide information on interrater agreement for a sample of 20% of task raters. These tables show, for each of the tasks, the percentage of agreement between two raters. The first column shows the task and the second column shows the number of responses that were double scored. DRC selects a sample of 20% of all responses scored, chosen at random during the operational scoring process. The next columns show the rates of agreement.

For Writing, with 0–6 as defined levels and the possibility of awarding a "plus" score between levels (e.g., 3, 3+, or 4 are all valid scores), scores that match or are contiguous (for example, if Rater 1 assigns a 3+ and Rater 2 assigns a score of 3, 3+, or 4) are categorized as agreement (%AG). Scores that are one whole score point apart (for example, if Rater 1 assigns a 3+ and Rater 2 assigns a score of 2+ or 4+) are categorized as adjacent (%AD). Otherwise, the raters are nonadjacent (%NA). Note that for Writing, interrater agreement is computed independently between ratings of keyboarded and handwritten responses.

For Speaking, the rating scale ranges from 0 to 4. If the two raters agree on the rating, an exact agreement is counted (%EX). If the two raters differ by one point, an adjacent agreement is counted (%AD). Otherwise, the raters are nonadjacent (%NA). Note that the Speaking tasks that target PL1—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 the Speaking scoring scale with only three possible score points (see Part 1, Sections 2.1.4 and 3.2.4 for more detail). As the range of possible score points is smaller for these tasks, the rater agreement tends to be higher.

WIDA stipulates a minimum interrater agreement rate of 70%. Tasks with interrater agreement rates between 70% and 74% are regarded as borderline.

For Writing, the lowest value for interrater agreement was 95%. For Speaking, the lowest value for interrater agreement was 74%.

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 S501 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	85,804	98	2	0
	2	83,988	98	2	0

Table 5.2.3.1.2

Interrater Agreement: Writ 1 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	11,852	98	2	0
	2	12,134	98	2	0

5.2.3.2 *Grades 2–3*

Table 5.2.3.2.1

Interrater Agreement: Writ 2-3 A S501 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	57,050	97	3	0
	2	56,086	98	2	0

Table 5.2.3.2.2

Interrater Agreement: Writ 2-3 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% AG	% AD	% NA
	1	122,046	98	2	0
	2	121,636	96	4	0

5.2.3.3 *Grades 4–5*

Table 5.2.3.3.1

Interrater Agreement: Writ 4-5 A S501 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	4,218	99	1	0
		KB	18,698	98	2	0
	2	HW	4,558	99	1	0
		KB	18,766	98	2	0

Table 5.2.3.3.2

Interrater Agreement: Writ 4-5 B/C S501 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	13,064	98	2	0
		KB	106,466	97	3	0
	2	HW	12,540	97	3	0
		KB	103,932	97	3	0

5.2.3.4 *Grades 6–8*

Table 5.2.3.4.1

Interrater Agreement: Writ 6-8 A S501 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	300	98	2	0
		KB	47,646	98	2	0
	2	HW	336	97	3	0
		KB	47,090	97	3	0

Table 5.2.3.4.2

Interrater Agreement: Writ 6-8 B/C S501 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	444	98	2	0
		KB	88,330	97	3	0
	2	HW	444	95	5	0
		KB	89,844	97	3	0

5.2.3.5 *Grades 9–12*

Table 5.2.3.5.1

Interrater Agreement: Writ 9-12 A S501 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	114	96	4	0
		KB	49,626	97	3	0
	2	HW	108	100	0	0
		KB	49,616	95	5	0

Table 5.2.3.5.2

Interrater Agreement: Writ 9-12 B/C S501 Online

Interrater Agreement	Task	Mode of Response	No. in Sample	% AG	% AD	% NA
	1	HW	132	97	3	0
		KB	87,650	98	2	0
	2	HW	122	97	3	0
		KB	90,394	97	3	0

5.2.4 **Speaking**

5.2.4.1 *Grade 1*

Table 5.2.4.1.1

Interrater Agreement: Spek 1 Pre-A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	5,296	98	2	0
	2	5,586	99	1	0
	3	5,674	98	2	0

Table 5.2.4.1.2

Interrater Agreement: Spek 1 A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	37,306	99	1	0
	2	37,306	85	14	0
	3	38,778	99	1	0
	4	38,778	86	13	0
	5	37,936	99	1	0
	6	37,936	88	12	0

Table 5.2.4.1.3

Interrater Agreement: Spek 1 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	52,516	85	15	0
	2	52,516	87	13	0
	3	50,934	85	14	0
	4	50,938	81	19	0
	5	51,290	85	15	0
	6	51,298	81	19	0

5.2.4.2 *Grades 2–3*

Table 5.2.4.2.1

Interrater Agreement: Spek 2-3 Pre-A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	11,310	98	2	0
	2	10,494	99	1	0
	3	10,744	98	2	0

Table 5.2.4.2.2

Interrater Agreement: Spek 2-3 A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	42,924	99	1	0
	2	42,914	85	15	0
	3	43,630	99	1	0
	4	43,536	88	12	0
	5	42,682	99	1	0
	6	42,682	82	17	1

Table 5.2.4.2.3

Interrater Agreement: Spek 2-3 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	123,744	83	17	0
	2	123,738	80	19	0
	3	123,510	84	15	0
	4	123,510	78	22	0
	5	125,588	80	20	0
	6	125,590	74	25	1

5.2.4.3 *Grades 4–5*

Table 5.2.4.3.1

Interrater Agreement: Spek 4-5 Pre-A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	4,486	99	1	0
	2	4,808	99	1	0
	3	4,076	99	1	0

Table 5.2.4.3.2

Interrater Agreement: Spek 4-5 A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	17,154	99	1	0
	2	17,154	89	11	0
	3	18,634	99	1	0
	4	18,636	90	10	0
	5	17,020	99	1	0
	6	17,020	83	17	0

Table 5.2.4.3.3

Interrater Agreement: Spek 4-5 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	128,524	80	19	0
	2	128,524	79	21	0
	3	131,308	84	16	0
	4	131,308	80	20	0
	5	129,620	81	19	0
	6	129,616	76	23	0

5.2.4.4 *Grades 6–8*

Table 5.2.4.4.1

Interrater Agreement: Spek 6-8 Pre-A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	5,690	99	1	0
	2	5,782	99	1	0
	3	5,588	99	1	0

Table 5.2.4.4.2

Interrater Agreement: Spek 6-8 A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	33,860	99	1	0
	2	33,860	87	13	0
	3	33,374	99	1	0
	4	33,372	86	14	0
	5	33,574	99	1	0
	6	33,574	89	11	0

Table 5.2.4.4.3

Interrater Agreement: Spek 6-8 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	119,790	82	18	0
	2	119,790	80	20	0
	3	118,810	80	19	1
	4	118,820	79	20	1
	5	113,058	83	17	1
	6	113,064	78	21	1

5.2.4.5 *Grades 9–12*

Table 5.2.4.5.1

Interrater Agreement: Spek 9-12 Pre-A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	12,622	99	1	0
	2	11,238	99	1	0
	3	12,078	99	1	0

Table 5.2.4.5.2

Interrater Agreement: Spek 9-12 A S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	71,846	99	1	0
	2	71,838	83	16	1
	3	69,314	99	1	0
	4	69,314	82	18	0
	5	71,558	100	0	0
	6	71,558	84	16	1

Table 5.2.4.5.3

Interrater Agreement: Spek 9-12 B/C S501 Online

Interrater Agreement	Task	No. in Sample	% EX	% AD	% NA
	1	85,016	76	23	0
	2	85,016	79	20	0
	3	84,826	77	23	1
	4	84,826	77	22	1
	5	84,772	83	17	0
	6	84,772	81	18	0

5.3 Conditional Standard Errors of Measurement at Cut Score

The tables in this section present information on the conditional standard errors of measurement (CSEM) at the most important points at which decisions are made about students based on performance on ACCESS—the cut points between language proficiency levels. Because the cut points depend on the grade level, information is provided for each grade level within a grade-level cluster.

Since the Listening and Reading tests are multistage adaptive tests, the CSEM will vary for the same scale score since students are routed to take different items; therefore, it is not possible to present a single value for the CSEM of the scale score that corresponds to each cut score. 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 score in the scale score metric (e.g., 305). The next columns present number of students and the minimum, maximum, mean, and standard deviation of the CSEM of all students at the cut scores. Note that there are some rare cases where there are no observed scale scores corresponding to the cut score values; therefore, these descriptive statistics cannot be provided.

For Writing and Speaking, the values are presented 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 cuts. For example, Tier A is intended for students at the lowest end of the language proficiency continuum. Optimally, Tier A forms should have the lowest CSEM of any tier at the 1/2 proficiency level cut and a relatively low CSEM at the 2/3 proficiency level cut. At the other end of the continuum, Tier B/C forms should optimally have a relatively low CSEM at the 4/5 proficiency level cut. These tables provide comparable information on how well the two tier forms are targeted to provide the most accurate measure in order to place their intended examinees into the language proficiency levels that they target. In the tables for Writing and Speaking, 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 score in the scale score metric (e.g., 305). In the last column(s), the corresponding CSEM is given for each cut score in the scale score metric for Writing and Speaking.

As a general rule, lower CSEM values around decision points are desirable. For the ACCESS population, CSEM values for the highest cut points are typically high. Students are exited from the ACCESS population upon gaining English language proficiency, and therefore these students are removed from the ACCESS population, resulting in smaller numbers of students at the highest cut points.

5.3.1 **Listening**

5.3.1.1 *Grade 1*

Table 5.3.1.1

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: List 1 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	1	236	N/A	N/A	N/A	N/A	N/A
2/3	1	259	118	16.84	17.35	16.99	0.23
3/4	1	291	4,035	16.84	17.86	16.84	0.09
4/5	1	303	1,170	16.84	17.86	16.90	0.22
5/6	1	327	7	18.37	18.37	18.37	0.00

5.3.1.2 *Grades 2–3*

Table 5.3.1.2

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: List 2-3 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	2	245	N/A	N/A	N/A	N/A	N/A
	3	262	83	20.92	20.92	20.92	0.00
2/3	2	283	16	18.88	18.88	18.88	0.00
	3	300	4,003	17.35	18.37	18.21	0.37
3/4	2	314	424	18.88	20.92	19.06	0.48
	3	331	1,476	18.37	20.41	18.54	0.39
4/5	2	330	1,106	18.88	22.45	19.80	0.61
	3	349	280	20.41	22.45	20.92	0.72
5/6	2	354	3	21.43	21.43	21.43	0.00
	3	374	126	26.02	26.02	26.02	0.00

5.3.1.3 *Grades 4–5*

Table 5.3.1.3

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: List 4-5 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	4	275	N/A	N/A	N/A	N/A	N/A
	5	285	1	18.37	18.37	18.37	0.00
2/3	4	313	1,277	16.84	16.84	16.84	0.00
	5	323	263	16.84	16.84	16.84	0.00
3/4	4	343	239	17.86	18.37	18.33	0.13
	5	354	444	18.37	19.39	18.37	0.05
4/5	4	363	284	18.37	18.88	18.72	0.23
	5	375	156	18.88	19.90	19.25	0.49
5/6	4	388	N/A	N/A	N/A	N/A	N/A
	5	401	2	19.39	19.90	19.64	0.36

5.3.1.4 *Grades 6–8*

Table 5.3.1.4

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: List 6-8 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	6	294	6	18.88	19.90	19.39	0.56
	7	302	52	19.39	19.39	19.39	0.00
	8	308	N/A	N/A	N/A	N/A	N/A
2/3	6	332	13	16.33	16.33	16.33	0.00
	7	340	32	16.84	17.86	17.16	0.48
	8	347	17	16.84	16.84	16.84	0.00
3/4	6	363	74	16.84	17.35	17.30	0.15
	7	370	277	16.33	17.35	17.13	0.36
	8	377	57	16.33	16.84	16.68	0.24
4/5	6	385	89	16.33	16.84	16.56	0.26
	7	394	2,668	16.84	17.35	17.28	0.17
	8	402	1,659	16.84	17.86	17.33	0.50
5/6	6	411	348	16.84	17.35	17.03	0.25
	7	420	5,266	16.84	16.84	16.84	0.00
	8	427	314	17.86	17.86	17.86	0.00

5.3.1.5 *Grades 9–12*

Table 5.3.1.5

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: List 9-12 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	9	314	38	20.41	20.41	20.41	0.00
	10	325	16	19.90	20.92	20.15	0.46
	11	335	15	20.41	20.41	20.41	0.00
	12	342	3	17.35	17.35	17.35	0.00
2/3	9	353	562	16.33	16.84	16.37	0.14
	10	358	85	16.33	16.84	16.34	0.08
	11	364	47	16.33	16.84	16.67	0.24
	12	368	5	16.33	16.33	16.33	0.00
3/4	9	383	459	16.33	16.84	16.81	0.10
	10	389	712	16.33	17.86	16.87	0.18
	11	394	279	16.84	17.86	16.90	0.25
	12	398	67	16.84	18.88	17.57	0.99
4/5	9	409	740	17.35	17.86	17.60	0.26
	10	415	1,198	17.35	18.37	18.27	0.30
	11	420	246	17.86	18.37	18.27	0.20
	12	426	201	18.37	18.88	18.52	0.23
5/6	9	434	231	18.37	18.37	18.37	0.00
	10	441	590	19.39	22.45	19.74	0.55
	11	447	140	21.94	21.94	21.94	0.00
	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 of Conditional Standard Error of Measurement at Cut Scores: Read 1 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	1	264	1,361	10.20	12.76	11.66	0.44
2/3	1	286	7,364	9.69	10.20	9.71	0.08
3/4	1	304	454	9.69	10.20	9.73	0.12
4/5	1	315	206	9.69	10.20	10.18	0.10
5/6	1	334	2,512	10.20	10.71	10.21	0.03

5.3.2.2 *Grades 2–3*

Table 5.3.2.2

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: Read 2-3 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	2	283	882	10.71	12.24	11.67	0.29
	3	297	99	10.20	11.22	10.84	0.29
2/3	2	307	2,679	10.20	10.71	10.29	0.19
	3	323	3,756	9.69	10.20	9.74	0.14
3/4	2	326	2,621	9.69	10.20	10.20	0.05
	3	342	797	9.69	10.71	10.09	0.28
4/5	2	337	16	10.20	10.20	10.20	0.00
	3	352	7,763	10.20	11.73	10.22	0.09
5/6	2	355	413	10.71	10.71	10.71	0.00
	3	370	7,041	11.22	12.24	11.23	0.03

5.3.2.3 *Grades 4–5*

Table 5.3.2.3

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: Read 4-5 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	4	307	475	10.71	12.24	11.75	0.24
	5	316	466	10.20	12.24	11.74	0.24
2/3	4	335	222	10.20	11.22	10.45	0.37
	5	345	1,005	10.20	10.71	10.26	0.16
3/4	4	354	6,206	10.20	10.71	10.70	0.08
	5	364	577	10.20	11.22	10.65	0.17
4/5	4	364	208	10.20	10.71	10.61	0.20
	5	373	522	10.20	11.22	10.58	0.23
5/6	4	382	28	10.71	11.22	10.90	0.25
	5	391	12	10.71	11.73	11.27	0.26

5.3.2.4 *Grades 6–8*

Table 5.3.2.4

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: Read 6-8 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	6	323	388	12.24	12.24	12.24	0.00
	7	329	255	11.73	11.73	11.73	0.00
	8	335	609	11.22	11.73	11.25	0.11
2/3	6	353	310	10.20	11.73	10.23	0.19
	7	360	1,249	9.69	11.22	9.85	0.30
	8	366	646	10.20	11.22	10.25	0.21
3/4	6	373	494	10.20	11.73	10.30	0.21
	7	380	1,167	10.20	11.73	10.35	0.24
	8	386	136	10.20	11.22	10.47	0.26
4/5	6	382	358	10.20	11.22	10.81	0.38
	7	389	2,072	10.20	11.22	10.75	0.15
	8	395	178	10.71	11.73	11.06	0.38
5/6	6	399	25	10.71	13.27	11.47	0.53
	7	406	46	11.73	11.73	11.73	0.00
	8	412	210	11.73	12.24	11.83	0.20

5.3.2.5 *Grades 9–12*

Table 5.3.2.5

Descriptive Statistics of Conditional Standard Error of Measurement at Cut Scores: Read 9-12 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	No. of Students	Min.	Max.	Mean	Std. Dev.
1/2	9	340	277	11.22	12.76	11.30	0.28
	10	344	28	11.22	11.73	11.72	0.10
	11	348	592	11.22	11.73	11.23	0.02
	12	352	147	11.22	12.24	11.84	0.30
2/3	9	372	692	10.20	10.71	10.20	0.02
	10	377	194	10.20	10.71	10.24	0.13
	11	382	527	10.20	11.22	10.24	0.14
	12	386	589	10.20	10.71	10.20	0.02
3/4	9	392	284	10.20	10.71	10.32	0.21
	10	397	247	10.20	10.71	10.49	0.25
	11	402	132	10.20	11.22	10.64	0.35
	12	407	111	10.20	12.24	10.90	0.37
4/5	9	401	221	10.20	11.22	10.50	0.27
	10	406	140	10.20	11.22	10.78	0.24
	11	410	268	10.20	11.73	10.73	0.17
	12	414	70	10.71	12.76	11.20	0.34
5/6	9	418	21	10.71	12.24	10.81	0.35
	10	423	26	10.71	12.24	11.54	0.38
	11	427	183	11.22	11.73	11.70	0.13
	12	432	N/A	N/A	N/A	N/A	N/A

5.3.3 **Writing**

5.3.3.1 *Grade 1*

Table 5.3.3.1

Conditional Standard Error of Measurement at Cut Scores: Writ 1 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	1	238	14.50	14.50
2/3	1	275	20.41	18.53
3/4	1	337	20.94	21.75
4/5	1	382	18.80	18.72
5/6	1	405	23.09	19.87

5.3.3.2 *Grades 2–3*

Table 5.3.3.2

Conditional Standard Error of Measurement at Cut Scores: Writ 2-3 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	2	242	14.78	16.38
	3	247	15.31	16.00
2/3	2	279	20.41	17.99
	3	283	20.68	18.53
3/4	2	341	20.94	21.48
	3	346	20.68	21.21
4/5	2	388	19.06	19.33
	3	394	19.87	19.60
5/6	2	411	23.90	20.94
	3	418	26.58	22.29

5.3.3.3 *Grades 4–5*

Table 5.3.3.3

Conditional Standard Error of Measurement at Cut Scores: Writ 4-5 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	4	266	14.50	26.85
	5	267	14.50	26.05
2/3	4	288	16.65	16.11
	5	293	17.72	15.04
3/4	4	351	21.75	20.94
	5	356	21.75	21.21
4/5	4	401	19.06	21.48
	5	407	18.80	20.94
5/6	4	425	19.60	19.60
	5	433	20.94	19.06

5.3.3.4 *Grades 6–8*

Table 5.3.3.4

Conditional Standard Error of Measurement at Cut Scores: Writ 6-8 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	6	268	14.77	14.77
	7	273	14.77	14.23
	8	281	15.31	14.23
2/3	6	298	17.76	16.92
	7	305	18.80	18.26
	8	311	19.87	19.33
3/4	6	361	21.75	21.75
	7	367	21.48	21.75
	8	372	21.21	21.48
4/5	6	413	18.80	18.80
	7	419	19.06	18.53
	8	424	19.33	18.53
5/6	6	441	22.02	20.46
	7	450	24.70	23.09
	8	459	28.46	26.31

5.3.3.5 *Grades 9–12*

Table 5.3.3.5

Conditional Standard Error of Measurement at Cut Scores: Writ 9-12 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	9	289	14.23	14.50
	10	298	15.04	14.77
	11	308	16.65	15.96
	12	318	18.53	17.72
2/3	9	319	18.80	17.83
	10	326	19.87	19.06
	11	335	20.68	20.14
	12	344	21.34	20.94
3/4	9	378	21.75	21.75
	10	385	21.48	21.48
	11	391	21.21	21.21
	12	398	20.68	20.94
4/5	9	430	18.53	18.80
	10	436	18.80	18.80
	11	441	19.06	18.80
	12	447	19.87	19.33
5/6	9	469	25.78	24.17
	10	479	30.61	28.46
	11	490	37.86	34.64
	12	501	46.45	42.43

5.3.4 **Speaking**

5.3.4.1 *Grade 1*

Table 5.3.4.1

Conditional Standard Error of Measurement at Cut Scores: Spek 1 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	1	205	21.06	15.50
2/3	1	261	28.37	19.89
3/4	1	311	24.28	17.45
4/5	1	361	28.37	19.60
5/6	1	403	46.50	30.42

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 Error of Measurement at Cut Scores: Spek 2-3 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	2	220	23.40	16.67
	3	234	25.74	17.55
2/3	2	273	27.79	19.30
	3	283	26.62	19.01
3/4	2	322	23.98	17.55
	3	332	24.28	17.55
4/5	2	374	33.05	22.23
	3	386	38.31	25.15
5/6	2	415	58.20	35.97
	3	425	68.44	41.82

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 Error of Measurement at Cut Scores: Spek 4-5 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	4	246	23.11	16.09
	5	258	24.28	16.38
2/3	4	293	27.20	18.72
	5	302	27.20	19.30
3/4	4	342	24.86	18.43
	5	350	24.57	17.84
4/5	4	397	29.25	18.72
	5	407	31.88	19.89
5/6	4	435	43.58	26.03
	5	443	48.55	28.66

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 Error of Measurement at Cut Scores: Spek 6-8 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	6	268	23.11	15.79
	7	277	24.57	16.38
	8	284	25.74	16.96
2/3	6	310	28.66	19.60
	7	317	28.37	19.89
	8	323	28.08	19.89
3/4	6	360	24.28	17.84
	7	369	23.98	17.55
	8	377	23.98	17.19
4/5	6	417	30.13	19.60
	7	425	33.05	20.77
	8	433	35.97	22.23
5/6	6	451	46.21	27.49
	7	457	50.31	29.83
	8	463	55.28	32.46

Note: Tier Pre-A is not presented as it is not possible for Tier Pre-A students to receive a proficiency level higher than 2.

5.3.4.5 *Grades 9–12*

Table 5.3.4.5

Conditional Standard Error of Measurement at Cut Scores: Spek 9-12 S501 Online

Proficiency Level Cut Point	Grade	Cut Score	SEM	
			Tier A	Tier B/C
1/2	9	290	25.74	17.26
	10	295	26.62	17.55
	11	299	27.20	18.13
	12	302	27.49	18.43
2/3	9	328	27.49	19.89
	10	333	27.20	19.60
	11	337	26.91	19.60
	12	340	26.32	19.30
3/4	9	385	24.28	17.26
	10	393	25.15	17.26
	11	400	25.81	17.55
	12	406	26.91	17.84
4/5	9	440	37.73	23.11
	10	446	40.95	24.86
	11	451	43.87	26.32
	12	455	46.21	27.49
5/6	9	468	55.86	32.76
	10	471	58.79	34.22
	11	474	61.42	35.68
	12	476	63.17	36.85

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 psychometric property of interest is how accurately and consistently ACCESS domain scores can classify students into WIDA proficiency categories determined by the 2016 ACCESS standard setting process (Cook & MacGregor, 2017). The accuracy and consistency of these classifications can be useful for test users to judge the utility of this information and to policy makers to make decisions about test design and score reporting (American Educational Research Association et al., 2014). The analyses utilize the methods outlined by Livingston and Lewis (1995) and Young and Yoon (1998), 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 the observed test scores would agree with those made on the basis of their true scores (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 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 taken by Livingston and Lewis (1995) and implemented here uses information about the reliability of the test, the cut scores, and the observed distribution of scores. Then, using a four-parameter beta distribution, the distribution of the true scores and of scores on a parallel form is modeled. The Livingston and Lewis procedure requires that the reliability estimate of the test form be provided in estimating the classification consistency and accuracy statistics. For Listening and Reading, the Rasch student reliability estimates by grade-level clusters were used in the procedure. Since the Writing and Speaking tests were tiered, it was necessary to produce a single reliability estimate across tiers for 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 the administered test and the true score distribution. For example, an overall accuracy of 0.774 means that 77% of students would be classified into the correct performance level across all six proficiency levels according to

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 of 0.664 means that 66% of students would be classified into the same performance level if two parallel forms were administered. Classification consistency values are always lower than the corresponding classification accuracy values, because in classification consistency, both of the classifications are subject to measurement error. In classification accuracy, only one of the classifications is based on a score that contains error.

Marginal Classification Accuracy and Consistency

Overall classification accuracy and consistency 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 at the specific cut point (e.g., at the PL 4 or PL 5 cut). The statistics that can address this question are **marginal classification accuracy and consistency** or classification accuracy and consistency indices at the cut score level. These two terms are used interchangeably in this report. From an accountability perspective, the most important information for test users and policy makers to examine is the marginal classification accuracy and consistency.

The **classification accuracy indices at the cut score** examine the percentage of students who are accurately placed above and below the cut score. A classification accuracy index at cut score 4/5 of 0.774 means that 77% of students would be classified in the same way if they were classified according to their observed score and their true score, either into the proficiency levels below the cut score (i.e., PL 1 to PL 4) or into the proficiency levels above the cut score (i.e., PL 5 to PL 6). The **classification consistency indices at the cut score** examine the percentage of students classified consistently above and below the cut score. A classification consistency index at cut score 4/5 of 0.664 means that 66% of students would be classified in the same way if two parallel forms were administered, either into the proficiency levels below the cut score (i.e., PL 1 to PL 4) or into the proficiency levels above the cut score (i.e., PL 5 to PL 6). Note that the accuracy and consistency are generally higher at the cut scores than over the proficiency levels, or the overall classification accuracy and consistency. This is because the accuracy and consistency indices at the cut examine the classification decisions at one cut point at a time while the overall accuracy and consistency statistics examine the classification decisions at all five ACCESS cut scores at the same time.

Classification accuracy and consistency indices are affected by the interaction of the number of proficiency cuts, the magnitude of the test reliability coefficient, measurement accuracy at the cut score, the distance between adjacent cuts, the location of the cut scores on the ability scale, and the proportion of students around a cut score (Lee, Hanson, & Brennan, 2002; Ercikan & Julian, 2002), and these factors are functions of the test design and most importantly the standard-setting decisions. The greater the number of proficiency levels, the lower the test reliability, the higher the measurement accuracy at the cut scores, the closer the two adjacent cut

scores, and the greater the proportion of students around a cut score, the lower the indices. Furthermore, the test reliability coefficient is affected by the numbers and types of items. For example, the test reliability estimate for the ACCESS Online Writing domain would be lower than similar tests with more items or tasks since it is estimated based on only two tasks.

For each test domain, we present three tables. The first provides the overall accuracy and the overall consistency for each grade level. The second provides the classification accuracy at the cut score for each grade level. The third provides the classification consistency at the cut score for each grade level. If the overall and marginal classification accuracy and consistency indices cannot be estimated because there are fewer than 200 students in the proficiency level, we collapsed the affected proficiency level category with the category below it and placed 'N/A' in the table for the affected proficiency level.

There has been very little guidance for the ideal or expected levels of decision consistency and accuracy needed for educational assessments since these statistics are affected by many different factors, as discussed earlier. We summarize the range of overall classification accuracy and consistency of domains across grades and highlight the grade level with the lowest classification accuracy and consistency for test users and policy makers. Since the overall accuracy and consistency statistics are a summary of the degree of classification accuracy and consistency across all proficiency level cut points, the marginal classification accuracy and consistency for these grades were further examined to identify the specific source(s) of low classification accuracy and consistency.

For Listening, as shown in Table 5.4.1.1, overall classification accuracy ranged from 0.567 to 0.806 and overall classification consistency ranged from 0.460 to 0.756. The lowest overall classification accuracy and consistency values were found for students in Grade 10.

For Reading, as shown in Table 5.4.2.1, overall classification accuracy ranged from 0.605 to 0.688 and overall classification consistency ranged from 0.502 to 0.594. The lowest overall classification accuracy and consistency values were found for students in Grade 4 for classification accuracy and Grade 2 for classification consistency.

For Writing, as shown in Table 5.4.3.1, overall classification accuracy ranged from 0.560 to 0.753 and overall classification consistency ranged from 0.484 to 0.636. The lowest overall classification accuracy and consistency values were found for students in Grade 5.

For Speaking, as shown in Table 5.4.4.1, overall classification accuracy ranged from 0.605 to 0.752 and overall classification consistency ranged from 0.514 to 0.651. The lowest overall classification accuracy and consistency values were found for students in Grade 5.

The results suggest that Grades 4 and 5 together had the lowest overall classification accuracy and consistency of the domains in three of the four domains (Reading, Writing, and Speaking). Grade 10 had the lowest overall classification accuracy and consistency in the Listening domain.

From an accountability perspective, the most important information for test users and policy makers to examine is the marginal classification accuracy and consistency. We summarize the range of the marginal classification accuracy and consistency of domains across grades and highlight the grade level with the lowest marginal classification accuracy and the lowest consistency by domain, for test users and policy makers.

For Listening, classification accuracy at the cut ranged from 0.869 to 0.988 (Table 5.4.1.2) and classification consistency at the cut ranged from 0.821 to 0.985 (Table 5.4.1.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 9 at the PL 3/PL 4 cut followed by Grade 10. Note that Grade 10 was also identified as having the lowest overall classification accuracy and consistency in the Listening domain. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal classification accuracy and consistency for Grades 9 and 10 Listening are still in the 80's.

For Reading, classification accuracy indices at the cut ranged from 0.877 to 0.922 (Table 5.4.2.2) and classification consistency at the cut ranged from 0.836 to 0.958 (Table 5.4.2.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 4 at the PL 4/PL 5 cut. Note that Grade 4 was also identified as having the lowest overall classification accuracy in the Reading domain. The low marginal classification accuracy and consistency at the PL 4/PL 5 cut appeared to have contributed to its low overall classification accuracy. However, it should be noted that the marginal classification accuracy and consistency for Grade 4 Reading are still in the 80's.

For Writing, classification accuracy indices at the cut ranged from 0.639 to 0.999 (Table 5.4.3.2), and classification consistency at the cut ranged from 0.581 to 0.998 (Table 5.4.3.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 4 at the PL 3/PL 4 cut followed by Grade 5. Note that Grade 5 was also identified as having the lowest overall classification accuracy and consistency in the Writing domain. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut appeared to have contributed to its low overall classification accuracy and consistency.

For Speaking, classification accuracy indices at the cut ranged from 0.768 to 0.998 (Table 5.4.4.2) and classification consistency at the cut ranged from 0.717 to 0.999 (Table 5.4.4.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 5 at the PL 3/PL 4 cut. Note that Grade 5 was also identified as having the lowest overall classification accuracy and consistency. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal classification accuracy and consistency for Grade 5 Speaking are still in the 70's.

The results from the overall and marginal classification accuracy and consistency statistics provided similar findings. In particular, Grades 4 and 5 together (Grades 4–5 cluster) had the lowest overall and marginal classification accuracy and consistency in three out of the four domains (Reading, Writing, and Speaking), and Grade 10 had the lowest overall and marginal classification accuracy in the Listening domain.

In addition, we observed that the lowest marginal classification accuracy and consistency of the domains occurred at the PL 3/PL 4 and PL 4/PL 5 cut points. This finding is consistent with previous research (Lee et al., 2000; Ercikan & Julian, 2002) in that classification accuracy and consistency at cut points in the middle of the proficiency level range are lower than those in the lower and upper ends. One possible reason might be that the cut scores for the proficiency level categories in the middle of the proficiency level range tend to be closer together as compared to those on the ends. The higher number of proficiency levels typically results in cut scores that are closer to each other than if a smaller number of proficiency levels were used. Classification accuracy and consistency are expected to vary for different ability levels due to variation in measurement accuracy. The further away the scores are from the cut scores, the smaller the classification errors would be or the more accurate the classification decisions would be. With a large number of proficiency levels, there are more students near the cut scores 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, and PL 3 and PL 4 occupy relatively narrow ranges on the ability scale as compared to other proficiency levels, the classification accuracy and consistency for the 3/4 and 4/5 cuts are lower than for other cuts.

Although there has been very little guidance for the ideal or expected levels of decision consistency and accuracy needed for educational assessments since these statistics are affected by many different factors, as discussed earlier, the range classification accuracy and consistency statistics for 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 ACCESS Online Writing consists of only two tasks, the test reliability estimate may be lower than similar writing tests with more items. The classification accuracy and consistency statistics derived using the Livingston and Lewis (1995) procedure are affected by the magnitude of the test reliability, which is lower when a test has fewer tasks. Also note that we do not expect the values estimated for ACCESS domains to be exactly the same as those computed in other programs, because testing programs differ in the student population, numbers of proficiency levels, test design, score distributions, and methods used to compute classification accuracy and consistency statistics. For example, ACCESS has a much larger and more diverse population and states, more proficiency levels, and a more complex test design than similar testing programs. Therefore, it is difficult to make an absolute comparison between the classification accuracy and consistency statistics for ACCESS domains with those from other testing programs.

5.4.1 **Listening**

Table 5.4.1.1

Overall Accuracy and Consistency of Classification Indices: List S501 Online

Grade	Accuracy	Consistency
1	0.680	0.614
2	0.623	0.544
3	0.630	0.551
4	0.806	0.756
5	0.791	0.738
6	0.657	0.567
7	0.624	0.535
8	0.617	0.531
9	0.573	0.465
10	0.567	0.460
11	0.570	0.464
12	0.571	0.462

Table 5.4.1.2

Classification Accuracy Indices at Cut Score Level: List S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.957	0.946	0.911	0.899	0.887
2	0.942	0.912	0.897	0.902	0.893
3	0.956	0.923	0.901	0.896	0.879
4	0.988	0.980	0.958	0.944	0.903
5	0.983	0.974	0.955	0.941	0.898
6	0.987	0.962	0.916	0.893	0.874
7	0.978	0.946	0.899	0.884	0.882
8	0.974	0.932	0.897	0.888	0.883
9	0.954	0.905	0.869	0.890	0.926
10	0.953	0.912	0.871	0.879	0.920
11	0.946	0.905	0.873	0.891	0.925
12	0.939	0.910	0.870	0.889	0.933

Table 5.4.1.3

Classification Consistency Indices at Cut Score Level: List S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.941	0.920	0.876	0.860	0.840
2	0.919	0.876	0.855	0.856	0.850
3	0.936	0.892	0.862	0.851	0.832
4	0.985	0.970	0.940	0.915	0.863
5	0.977	0.962	0.935	0.910	0.857
6	0.983	0.944	0.883	0.848	0.825
7	0.970	0.920	0.861	0.836	0.836
8	0.961	0.903	0.857	0.840	0.837
9	0.936	0.864	0.821	0.844	0.895
10	0.934	0.873	0.823	0.832	0.885
11	0.923	0.865	0.825	0.846	0.895
12	0.915	0.868	0.821	0.844	0.904

5.4.2 **Reading**

Table 5.4.2.1

Overall Accuracy and Consistency of Classification Indices: Read S501 Online

Grade	Accuracy	Consistency
1	0.619	0.513
2	0.611	0.502
3	0.611	0.508
4	0.605	0.504
5	0.607	0.507
6	0.685	0.590
7	0.688	0.594
8	0.681	0.591
9	0.683	0.590
10	0.666	0.571
11	0.656	0.560
12	0.661	0.565

Table 5.4.2.2

Classification Accuracy Indices at Cut Score Level: Read S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.886	0.883	0.914	0.943	0.969
2	0.943	0.898	0.890	0.907	0.951
3	0.928	0.897	0.906	0.906	0.936
4	0.959	0.922	0.891	0.877	0.921
5	0.948	0.909	0.889	0.888	0.927
6	0.928	0.903	0.920	0.939	0.972
7	0.920	0.905	0.929	0.941	0.965
8	0.920	0.905	0.926	0.939	0.957
9	0.916	0.918	0.932	0.933	0.954
10	0.926	0.917	0.923	0.921	0.943
11	0.936	0.915	0.913	0.910	0.938
12	0.931	0.918	0.911	0.911	0.948

Table 5.4.2.3

Classification Consistency Indices at Cut Score Level: Read S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.841	0.838	0.881	0.916	0.955
2	0.918	0.857	0.848	0.870	0.930
3	0.899	0.856	0.864	0.870	0.909
4	0.943	0.888	0.848	0.836	0.885
5	0.926	0.870	0.846	0.847	0.895
6	0.900	0.866	0.886	0.912	0.958
7	0.888	0.868	0.897	0.915	0.950
8	0.888	0.870	0.893	0.911	0.940
9	0.882	0.885	0.902	0.908	0.934
10	0.897	0.882	0.890	0.892	0.920
11	0.911	0.880	0.876	0.878	0.912
12	0.904	0.883	0.878	0.881	0.924

5.4.3 **Writing**

Table 5.4.3.1

Overall Accuracy and Consistency of Classification Indices: Writ S501 Online

Grade	Accuracy	Consistency
1	0.581	0.541
2	0.753	0.636
3	0.738	0.616
4	0.581	0.499
5	0.560	0.484
6	0.719	0.606
7	0.617	0.531
8	0.690	0.559
9	0.635	0.534
10	0.702	0.586
11	0.663	0.555
12	0.659	0.560

Table 5.4.3.2

Classification Accuracy Indices at Cut Score Level: Writ S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.915	0.665	0.993	N/A	N/A
2	0.959	0.844	0.947	N/A	N/A
3	0.977	0.931	0.829	0.999	N/A
4	0.982	0.957	0.639	0.978	0.995
5	0.981	0.960	0.650	0.958	0.997
6	0.960	0.883	0.871	N/A	N/A
7	0.945	0.850	0.816	N/A	N/A
8	0.932	0.852	0.894	N/A	N/A
9	0.932	0.881	0.815	0.997	N/A
10	0.948	0.885	0.865	0.995	N/A
11	0.939	0.861	0.860	0.994	N/A
12	0.933	0.870	0.845	N/A	N/A

Table 5.4.3.3

Classification Consistency Indices at Cut Score Level: Writ S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.869	0.652	0.993	N/A	N/A
2	0.938	0.790	0.885	N/A	N/A
3	0.965	0.903	0.736	0.998	N/A
4	0.973	0.941	0.581	0.967	0.994
5	0.972	0.944	0.582	0.940	0.996
6	0.936	0.832	0.812	N/A	N/A
7	0.917	0.799	0.783	N/A	N/A
8	0.898	0.790	0.825	N/A	N/A
9	0.903	0.830	0.766	0.993	N/A
10	0.922	0.832	0.806	0.993	N/A
11	0.908	0.811	0.809	0.992	N/A
12	0.900	0.812	0.805	N/A	N/A

5.4.4 **Speaking**

Table 5.4.4.1

Overall Accuracy and Consistency of Classification Indices: Spek S501 Online

Grade	Accuracy	Consistency
1	0.649	0.554
2	0.676	0.560
3	0.693	0.556
4	0.633	0.533
5	0.605	0.514
6	0.674	0.566
7	0.700	0.578
8	0.655	0.560
9	0.741	0.646
10	0.752	0.651
11	0.707	0.621
12	0.700	0.613

Table 5.4.4.2

Classification Accuracy Indices at Cut Score Level: Spek S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.935	0.864	0.852	0.989	N/A
2	0.943	0.826	0.906	0.994	N/A
3	0.954	0.845	0.893	0.994	N/A
4	0.959	0.884	0.807	0.987	0.998
5	0.955	0.888	0.768	0.984	N/A
6	0.946	0.862	0.866	0.996	N/A
7	0.940	0.857	0.899	0.998	N/A
8	0.925	0.853	0.870	0.995	N/A
9	0.909	0.862	0.961	N/A	N/A
10	0.922	0.856	0.968	N/A	N/A
11	0.916	0.826	0.959	N/A	N/A
12	0.918	0.796	0.979	N/A	N/A

Table 5.4.4.3

Classification Consistency Indices at Cut Score Level: Spek S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.904	0.811	0.820	0.988	N/A
2	0.915	0.764	0.852	0.994	N/A
3	0.932	0.777	0.814	0.994	N/A
4	0.938	0.843	0.743	0.985	0.999
5	0.933	0.843	0.717	0.982	N/A
6	0.920	0.813	0.812	0.996	N/A
7	0.912	0.798	0.837	0.997	N/A
8	0.894	0.794	0.829	0.995	N/A
9	0.872	0.802	0.934	N/A	N/A
10	0.887	0.790	0.943	N/A	N/A
11	0.880	0.768	0.944	N/A	N/A
12	0.881	0.737	0.967	N/A	N/A

5.5 Reliability of Composite Scores

The reliability of ACCESS composites evaluates the consistency of the composite scores of the students 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 attributes, even though items within the domain are assumed to measure similar attributes, a traditional internal consistency statistic such as Cronbach alpha is not appropriate, as such statistics were developed assuming items in a test measure similar attributes. It is more appropriate to report stratified alpha (Feldt & Brennan, 1989), derived to measure consistency in students' scores when the total score consists of heterogeneous parts. Stratified alpha is a weighted average of coefficient alphas for item sets with different maximum score points or "strata." Stratified alpha is a reliability estimate computed by dividing the test into parts (strata), computing Cronbach's alpha separately for each part, and using the results to estimate a reliability coefficient for the total score.

In computing the stratified Cronbach's alpha for ACCESS composites, each domain that makes up a composite is treated as a strata. For example, in computing stratified Cronbach's alpha for Literacy, two strata (Reading and Writing) are entered into the computation. The stratified Cronbach's alpha is interpreted like other traditional internal consistency statistics such as Cronbach's coefficient alpha. Like Cronbach's alpha, stratified Cronbach's alpha is an estimate of the proportion of the total variance of the composite that can be explained by the variance of the true score.

Because of the differential weights applied to the ACCESS domains that contribute to the composites, the stratified Cronbach's alpha coefficient is weighted by the contribution of each domain score into the composite (Rudner, 2001; Kamata, Turhan, & Darandari, 2003; Kane & Case, 2004). 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 = number of components j
w_j = weight of component j
σ_j^2 = variance of component j
σ_c^2 = variance of composite
ρ_j = reliability coefficient of component j

The tables below express the stratified Cronbach's alpha for each of the composites. The first table for each composite provides stratified Cronbach's alpha for all students. The second table for each composite provides the same information for the population of female students and the population of male students. The third table provides information by ethnicity, for Hispanic and

non-Hispanic students, and the fourth table provides information for the population of students who have an individualized education plan.

Each table is divided by grade-level cluster. Tables first include the input values used to compute Cronbach's alpha. The table lists the number of components for each composite and their weight. (Detail on how the composites are computed is provided in the introduction to Chapter 3.)

For the Listening and Reading domain components, the reliability coefficient is the Rasch student reliability coefficient, provided in Section 5.1.

For 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, values for Cronbach's alpha for each of the tiers in the grade-level cluster (provided in Section 5.1) are weighted by the number of students who were administered the tier form, and a weighted average is expressed in the tables.

For each relevant domain component, we provide the variance of the scale score. We also provide the variance of the composite scale score. The variances of domains and composites are computed for students who had valid results in all four domains.

Finally, the table presents the computed stratified Cronbach's alpha value for the composite, by grade-level cluster.

The stratified Cronbach's alpha, presented in the tables in this section, was also used to produce the *Accuracy and Consistency* classification tables of the composites (Section 5.7).

The stratified Cronbach's alpha of the Oral composite computed for all students ranged from 0.89 to 0.91. The stratified Cronbach's alpha ranged from 0.89 to 0.91 for male students; 0.88 to 0.90 for female students; 0.89 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 alpha of the Literacy composite computed for all students ranged from 0.86 to 0.88. The stratified Cronbach's alpha of the Literacy composite ranged from 0.87 to 0.89 for male students; 0.85 to 0.88 for female students; 0.86 to 0.88 for Hispanic students; 0.86 to 0.88 for non-Hispanic students; and 0.85 to 0.88 for students with an IEP.

The stratified Cronbach's alpha of the Comprehension composite computed for all students ranged from 0.91 to 0.94. The stratified Cronbach's alpha of the Comprehension composite ranged from 0.91 to 0.94 for male students; 0.91 to 0.93 for female students; 0.90 to 0.94 for Hispanic students; 0.91 to 0.94 for non-Hispanic students; and 0.89 to 0.91 for students with an IEP.

The stratified Cronbach's alpha of the Overall composite computed for all students ranged from 0.92 to 0.93. The stratified Cronbach's alpha of the Overall composite ranged from 0.92 to 0.94 for male students; 0.91 to 0.93 for female students; 0.92 to 0.93 for Hispanic students; 0.92 to 0.93 for non-Hispanic students; and 0.91 to 0.93 for students with an IEP.

5.5.1 **Oral**

Table 5.5.1.1

Reliability of Composite: Oral S501 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.50	3052.023	0.860
	Speaking	0.50	2957.389	0.827
	Oral		2281.629	0.897
2-3	Listening	0.50	3819.106	0.860
	Speaking	0.50	2598.528	0.802
	Oral		2561.371	0.898
4-5	Listening	0.50	2741.884	0.820
	Speaking	0.50	2593.505	0.815
	Oral		2151.566	0.887
6-8	Listening	0.50	2355.885	0.850
	Speaking	0.50	2685.628	0.826
	Oral		2002.910	0.897
9-12	Listening	0.50	2340.899	0.850
	Speaking	0.50	3333.456	0.850
	Oral		2238.480	0.905

Table 5.5.1.2

Reliability of Composite: Oral S501 Online by Gender

			Female		Male	
Cluster	Component	Weight	Variance	Reliability	Variance	Reliability
1	Listening	0.50	2923.405	0.860	3147.854	0.870
	Speaking	0.50	2943.900	0.829	2895.178	0.822
	Oral		2209.807	0.897	2303.994	0.900
2-3	Listening	0.50	3644.587	0.850	3969.540	0.860
	Speaking	0.50	2570.618	0.801	2570.930	0.800
	Oral		2469.930	0.893	2619.719	0.898
4-5	Listening	0.50	2580.870	0.810	2867.440	0.830
	Speaking	0.50	2527.785	0.811	2628.969	0.818
	Oral		2042.497	0.881	2233.508	0.892
6-8	Listening	0.50	2239.764	0.850	2449.786	0.860
	Speaking	0.50	2729.072	0.827	2652.240	0.825
	Oral		1965.845	0.897	2037.134	0.901
9-12	Listening	0.50	2229.905	0.840	2429.286	0.850
	Speaking	0.50	3294.098	0.843	3362.067	0.854
	Oral		2179.874	0.900	2287.522	0.907

Table 5.5.1.3

Reliability of Composite: Oral S501 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.50	2989.632	0.870	3029.220	0.850
	Speaking	0.50	2983.200	0.828	2744.345	0.819
	Oral		2266.039	0.901	2161.420	0.890
2-3	Listening	0.50	3811.076	0.870	3595.736	0.840
	Speaking	0.50	2687.418	0.807	2272.858	0.786
	Oral		2602.118	0.903	2288.124	0.884
4-5	Listening	0.50	2730.558	0.820	2460.678	0.790
	Speaking	0.50	2611.717	0.816	2266.685	0.805
	Oral		2149.731	0.887	1860.143	0.871
6-8	Listening	0.50	2321.228	0.850	2226.938	0.840
	Speaking	0.50	2722.041	0.828	2222.291	0.805
	Oral		1995.228	0.898	1737.664	0.886
9-12	Listening	0.50	2312.369	0.850	2231.305	0.840
	Speaking	0.50	3417.723	0.856	2743.741	0.820
	Oral		2255.204	0.907	1922.228	0.889

Table 5.5.1.4

Reliability of Composite: Oral S501 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.50	3283.451	0.890
	Speaking	0.50	3174.301	0.828
	Oral		2455.703	0.908
2-3	Listening	0.50	3723.605	0.880
	Speaking	0.50	2833.084	0.800
	Oral		2615.300	0.903
4-5	Listening	0.50	2275.433	0.810
	Speaking	0.50	2243.519	0.805
	Oral		1694.890	0.872
6-8	Listening	0.50	1790.199	0.820
	Speaking	0.50	2175.847	0.819
	Oral		1446.084	0.876
9-12	Listening	0.50	1743.181	0.810
	Speaking	0.50	3064.962	0.862
	Oral		1737.705	0.892

5.5.2 **Literacy**

Table 5.5.2.1

Reliability of Composite: Litr S501 Online

Cluster	Component	Weight	Variance	Reliability
1	Reading	0.50	1045.886	0.880
	Writing	0.50	1627.098	0.771
	Literacy		986.020	0.874
2-3	Reading	0.50	1049.662	0.880
	Writing	0.50	1926.363	0.740
	Literacy		1184.131	0.868
4-5	Reading	0.50	1158.966	0.890
	Writing	0.50	2086.405	0.711
	Literacy		1306.273	0.860
6-8	Reading	0.50	1463.467	0.910
	Writing	0.50	1461.014	0.659
	Literacy		1225.985	0.871
9-12	Reading	0.50	1539.119	0.910
	Writing	0.50	1555.005	0.715
	Literacy		1240.919	0.883

Table 5.5.2.2

Reliability of Composite: Litr S501 Online by Gender

			Female		Male	
Cluster	Component	Weight	Variance	Reliability	Variance	Reliability
1	Reading	0.50	1062.035	0.880	1031.029	0.880
	Writing	0.50	1470.140	0.750	1723.764	0.781
	Literacy		945.681	0.869	1005.896	0.875
2-3	Reading	0.50	1014.967	0.880	1075.369	0.880
	Writing	0.50	1799.148	0.712	1982.597	0.752
	Literacy		1126.969	0.858	1211.111	0.872
4-5	Reading	0.50	1098.982	0.880	1203.982	0.890
	Writing	0.50	1886.556	0.679	2205.588	0.726
	Literacy		1209.860	0.848	1365.416	0.865
6-8	Reading	0.50	1401.638	0.910	1506.956	0.910
	Writing	0.50	1396.784	0.611	1486.471	0.690
	Literacy		1167.412	0.857	1258.191	0.882
9-12	Reading	0.50	1490.630	0.910	1567.099	0.920
	Writing	0.50	1486.651	0.694	1579.991	0.725
	Literacy		1190.505	0.876	1262.758	0.889

Table 5.5.2.3

Reliability of Composite: Litr S501 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Reading	0.50	874.135	0.860	1257.732	0.900
	Writing	0.50	1615.698	0.783	1456.952	0.730
	Literacy		875.410	0.865	1053.781	0.877
2-3	Reading	0.50	987.500	0.870	1103.331	0.880
	Writing	0.50	2029.816	0.749	1528.557	0.706
	Literacy		1187.590	0.866	1060.273	0.863
4-5	Reading	0.50	1109.437	0.880	1210.038	0.890
	Writing	0.50	2116.222	0.712	1737.695	0.699
	Literacy		1293.387	0.857	1191.447	0.863
6-8	Reading	0.50	1388.778	0.910	1541.389	0.920
	Writing	0.50	1461.468	0.670	1307.065	0.616
	Literacy		1196.005	0.873	1177.529	0.867
9-12	Reading	0.50	1496.863	0.910	1499.003	0.910
	Writing	0.50	1593.810	0.723	1322.319	0.684
	Literacy		1244.597	0.884	1096.936	0.874

Table 5.5.2.4

Reliability of Composite: Litr S501 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Reading	0.50	744.598	0.830
	Writing	0.50	2048.386	0.821
	Literacy		917.180	0.865
2-3	Reading	0.50	879.508	0.850
	Writing	0.50	2122.908	0.812
	Literacy		1130.760	0.883
4-5	Reading	0.50	990.665	0.870
	Writing	0.50	1922.909	0.752
	Literacy		1120.094	0.865
6-8	Reading	0.50	1102.170	0.880
	Writing	0.50	1035.114	0.690
	Literacy		851.816	0.867
9-12	Reading	0.50	1023.067	0.880
	Writing	0.50	1135.304	0.705
	Literacy		759.198	0.849

5.5.3 **Comprehension**

Table 5.5.3.1

Reliability of Composite: Cphn S501 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.30	3052.023	0.860
	Reading	0.70	1045.886	0.880
	Comprehension		1104.182	0.909
2-3	Listening	0.30	3819.106	0.860
	Reading	0.70	1049.662	0.880
	Comprehension		1363.293	0.919
4-5	Listening	0.30	2741.884	0.820
	Reading	0.70	1158.966	0.890
	Comprehension		1314.672	0.919
6-8	Listening	0.30	2355.885	0.850
	Reading	0.70	1463.467	0.910
	Comprehension		1461.612	0.934
9-12	Listening	0.30	2340.899	0.850
	Reading	0.70	1539.119	0.910
	Comprehension		1539.512	0.935

Table 5.5.3.2

Reliability of Composite: Cphn S501 Online by Gender

			Female		Male	
Cluster	Component	Weight	Variance	Reliability	Variance	Reliability
1	Listening	0.30	2923.405	0.860	3147.854	0.870
	Reading	0.70	1062.035	0.880	1031.029	0.880
	Comprehension		1098.692	0.910	1105.566	0.912
2-3	Listening	0.30	3644.587	0.850	3969.540	0.860
	Reading	0.70	1014.967	0.880	1075.369	0.880
	Comprehension		1313.635	0.917	1404.233	0.919
4-5	Listening	0.30	2580.870	0.810	2867.440	0.830
	Reading	0.70	1098.982	0.880	1203.982	0.890
	Comprehension		1242.901	0.913	1371.677	0.921
6-8	Listening	0.30	2239.764	0.850	2449.786	0.860
	Reading	0.70	1401.638	0.910	1506.956	0.910
	Comprehension		1402.454	0.934	1510.168	0.936
9-12	Listening	0.30	2229.905	0.840	2429.286	0.850
	Reading	0.70	1490.630	0.910	1567.099	0.920
	Comprehension		1490.069	0.934	1574.935	0.940

Table 5.5.3.3

Reliability of Composite: Cphn S501 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.30	2989.632	0.870	3029.220	0.850
	Reading	0.70	874.135	0.860	1257.732	0.900
	Comprehension		948.538	0.900	1289.014	0.920
2-3	Listening	0.30	3811.076	0.870	3595.736	0.840
	Reading	0.70	987.500	0.870	1103.331	0.880
	Comprehension		1300.874	0.917	1383.093	0.916
4-5	Listening	0.30	2730.558	0.820	2460.678	0.790
	Reading	0.70	1109.437	0.880	1210.038	0.890
	Comprehension		1271.384	0.914	1303.061	0.914
6-8	Listening	0.30	2321.228	0.850	2226.938	0.840
	Reading	0.70	1388.778	0.910	1541.389	0.920
	Comprehension		1395.513	0.934	1493.719	0.938
9-12	Listening	0.30	2312.369	0.850	2231.305	0.840
	Reading	0.70	1496.863	0.910	1499.003	0.910
	Comprehension		1499.655	0.935	1484.683	0.934

Table 5.5.3.4

Reliability of Composite: Cphn S501 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.30	3283.451	0.890
	Reading	0.70	744.598	0.830
	Comprehension		845.014	0.888
2-3	Listening	0.30	3723.605	0.880
	Reading	0.70	879.508	0.850
	Comprehension		1139.965	0.908
4-5	Listening	0.30	2275.433	0.810
	Reading	0.70	990.665	0.870
	Comprehension		1040.898	0.902
6-8	Listening	0.30	1790.199	0.820
	Reading	0.70	1102.170	0.880
	Comprehension		1027.336	0.909
9-12	Listening	0.30	1743.181	0.810
	Reading	0.70	1023.067	0.880
	Comprehension		983.618	0.909

5.5.4 **Overall**

Table 5.5.4.1

Reliability of Composite: Over S501 Online

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.15	3052.023	0.860
	Reading	0.35	1045.886	0.880
	Writing	0.35	1627.098	0.771
	Speaking	0.15	2957.389	0.827
	Overall Composite		1046.081	0.921
2-3	Listening	0.15	3819.106	0.860
	Reading	0.35	1049.662	0.880
	Writing	0.35	1926.363	0.740
	Speaking	0.15	2598.528	0.802
	Overall Composite		1350.482	0.926
4-5	Listening	0.15	2741.884	0.820
	Reading	0.35	1158.966	0.890
	Writing	0.35	2086.405	0.711
	Speaking	0.15	2593.505	0.815
	Overall Composite		1366.431	0.918
6-8	Listening	0.15	2355.885	0.850
	Reading	0.35	1463.467	0.910
	Writing	0.35	1461.014	0.659
	Speaking	0.15	2685.628	0.826
	Overall Composite		1273.693	0.925
9-12	Listening	0.15	2340.899	0.850
	Reading	0.35	1539.119	0.910
	Writing	0.35	1555.005	0.715
	Speaking	0.15	3333.456	0.850
	Overall Composite		1350.187	0.933

Table 5.5.4.2

Reliability of Composite: Over S501 Online by Gender

Cluster	Component	Weight	Female		Male	
			Variance	Reliability	Variance	Reliability
1	Listening	0.15	2923.405	0.860	3147.854	0.870
	Reading	0.35	1062.035	0.880	1031.029	0.880
	Writing	0.35	1470.140	0.750	1723.764	0.781
	Speaking	0.15	2943.900	0.829	2895.178	0.822
	Overall Composite		1003.144	0.919	1063.196	0.923
2-3	Listening	0.15	3644.587	0.850	3969.540	0.860
	Reading	0.35	1014.967	0.880	1075.369	0.880
	Writing	0.35	1799.148	0.712	1982.597	0.752
	Speaking	0.15	2570.618	0.801	2570.930	0.800
	Overall Composite		1289.772	0.921	1381.612	0.928
4-5	Listening	0.15	2580.870	0.810	2867.440	0.830
	Reading	0.35	1098.982	0.880	1203.982	0.890
	Writing	0.35	1886.556	0.679	2205.588	0.726
	Speaking	0.15	2527.785	0.811	2628.969	0.818
	Overall Composite		1272.650	0.912	1428.624	0.922
6-8	Listening	0.15	2239.764	0.850	2449.786	0.860
	Reading	0.35	1401.638	0.910	1506.956	0.910
	Writing	0.35	1396.784	0.611	1486.471	0.690
	Speaking	0.15	2729.072	0.827	2652.240	0.825
	Overall Composite		1226.845	0.918	1305.005	0.930
9-12	Listening	0.15	2229.905	0.840	2429.286	0.850
	Reading	0.35	1490.630	0.910	1567.099	0.920
	Writing	0.35	1486.651	0.694	1579.991	0.725
	Speaking	0.15	3294.098	0.843	3362.067	0.854
	Overall Composite		1304.215	0.930	1376.798	0.936

Table 5.5.4.3

Reliability of Composite: Over S501 Online by Ethnicity

Cluster	Component	Weight	Hispanic		Other	
			Variance	Reliability	Variance	Reliability
1	Listening	0.15	2989.632	0.870	3029.220	0.850
	Reading	0.35	874.135	0.860	1257.732	0.900
	Writing	0.35	1615.698	0.783	1456.952	0.730
	Speaking	0.15	2983.200	0.828	2744.345	0.819
	Overall Composite		953.991	0.918	1084.728	0.922
2-3	Listening	0.15	3811.076	0.870	3595.736	0.840
	Reading	0.35	987.500	0.870	1103.331	0.880
	Writing	0.35	2029.816	0.749	1528.557	0.706
	Speaking	0.15	2687.418	0.807	2272.858	0.786
	Overall Composite		1357.225	0.926	1202.725	0.921
4-5	Listening	0.15	2730.558	0.820	2460.678	0.790
	Reading	0.35	1109.437	0.880	1210.038	0.890
	Writing	0.35	2116.222	0.712	1737.695	0.699
	Speaking	0.15	2611.717	0.816	2266.685	0.805
	Overall Composite		1354.858	0.917	1210.182	0.916
6-8	Listening	0.15	2321.228	0.850	2226.938	0.840
	Reading	0.35	1388.778	0.910	1541.389	0.920
	Writing	0.35	1461.468	0.670	1307.065	0.616
	Speaking	0.15	2722.041	0.828	2222.291	0.805
	Overall Composite		1246.555	0.926	1177.007	0.920
9-12	Listening	0.15	2312.369	0.850	2231.305	0.840
	Reading	0.35	1496.863	0.910	1499.003	0.910
	Writing	0.35	1593.810	0.723	1322.319	0.684
	Speaking	0.15	3417.723	0.856	2743.741	0.820
	Overall Composite		1355.507	0.934	1165.176	0.925

Table 5.5.4.4

Reliability of Composite: Over S501 Online by IEP Status

Cluster	Component	Weight	Variance	Reliability
1	Listening	0.15	3283.451	0.890
	Reading	0.35	744.598	0.830
	Writing	0.35	2048.386	0.821
	Speaking	0.15	3174.301	0.828
	Overall Composite		971.616	0.917
2-3	Listening	0.15	3723.605	0.880
	Reading	0.35	879.508	0.850
	Writing	0.35	2122.908	0.812
	Speaking	0.15	2833.084	0.800
	Overall Composite		1264.870	0.931
4-5	Listening	0.15	2275.433	0.810
	Reading	0.35	990.665	0.870
	Writing	0.35	1922.909	0.752
	Speaking	0.15	2243.519	0.805
	Overall Composite		1073.086	0.913
6-8	Listening	0.15	1790.199	0.820
	Reading	0.35	1102.170	0.880
	Writing	0.35	1035.114	0.690
	Speaking	0.15	2175.847	0.819
	Overall Composite		831.840	0.914
9-12	Listening	0.15	1743.181	0.810
	Reading	0.35	1023.067	0.880
	Writing	0.35	1135.304	0.705
	Speaking	0.15	3064.962	0.862
	Overall Composite		827.202	0.912

5.6 CSEM for Composites

Conditional standard errors of measurement (CSEMs) for the four ACCESS composites provide test users a benchmark of how free the composite scale score is from measurement errors at the various points of the composites. Due to the differential weights applied to different ACCESS domains (see the introduction to Section 3 for weighting conventions), we estimate the CSEMs using a procedure based on item response theory (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 CSEM of a composite consisting of subtests. The basic premise of this procedure is that the student-level CSEM for a weighted composite can be estimated empirically using the IRT-based CSEMs for each student on the subtests and the weights associated with the subtests. We used this method to estimate the CSEM for ACCESS composites by treating the ACCESS domains as subtests.

We use a three-step process to derive the CSEM for ACCESS composites. We conduct the derivation by grade and composite to obtain a unique CSEM for each composite score by grade. Since this procedure replies on empirical student data, which are subject to year-to-year fluctuation, we use all population student data from the previous ACCESS series in the derivation to obtain more stable estimates than using only data from a single series.

Step 1. Since we calibrated ACCESS domains separately, measurement errors associated with each of the ACCESS domains, as expressed in the conditional errors of measurement, are independent of each other. Therefore, the CSEM for a student with composite score *x*, SEM_x, can be estimated using the equation derived by Price et al. (2006):

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

Where SEM_i^2 is the student's IRT-based score error variance or student's squared CSEM in ACCESS domain *i* and W_i is the weight applied to domain *i*, for *i*=1,…,*k*.

Step 2. Due to the differential weights applied to different ACCESS domains, two students with the same sum of weighted domain score, or composite, may obtain different CSEMs; therefore, we took an additional step to obtain a unique value for each composite score. Specifically, we estimated the expected value of the CSEM functions for a composite score using a regression approach, and we reported this expected value as the CSEM for that composite score.

Step 3. A linear smoothing procedure was applied to derive the CSEMs for composite scores that were not observed in the data.

The figures in this section show graphically the CSEMs for various composite scores by grade level. Figures show the relationship between the students' composite scores on the horizontal axis and conditional measurement errors on the vertical axis. Each point in the figures represents a student in the dataset, expressing both the student's CSEM and that student's scale score for the given composite score. We do not plot values for students who received the lowest possible scores on any ACCESS domains, as it is not possible to compute accurately the conditional

measurement errors for these students. For grade-level clusters with multiple grades, we use different colors in the figures to represent students in different grades.

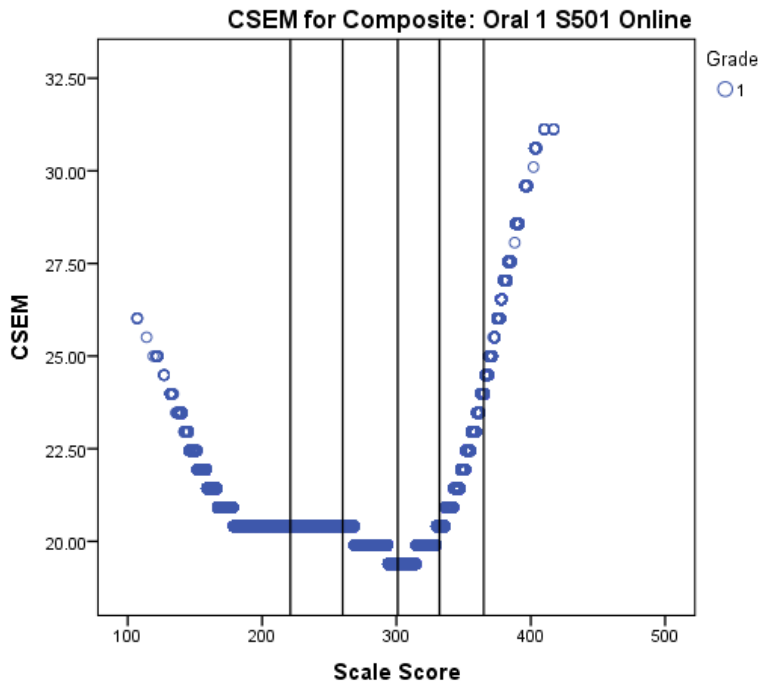
Five vertical lines in the figure indicate the five ACCESS cut scores for the highest grade in the grade-level cluster for the test form, dividing the figure into six sections for each of the WIDA proficiency levels (1–6) for the composites.

Low CSEM values indicate less measurement error or more accuracy in measurement. The general trend in these figures shows that the CSEMs are lower and fairly constant in the middle of the score range and higher and more variable for extreme low and high scores, as expected. As noted elsewhere in this report, students are exited from the ACCESS population upon gaining English language proficiency, and therefore these students are removed from the ACCESS population, resulting in smaller numbers of students at the highest cut points.

5.6.1 **Oral**

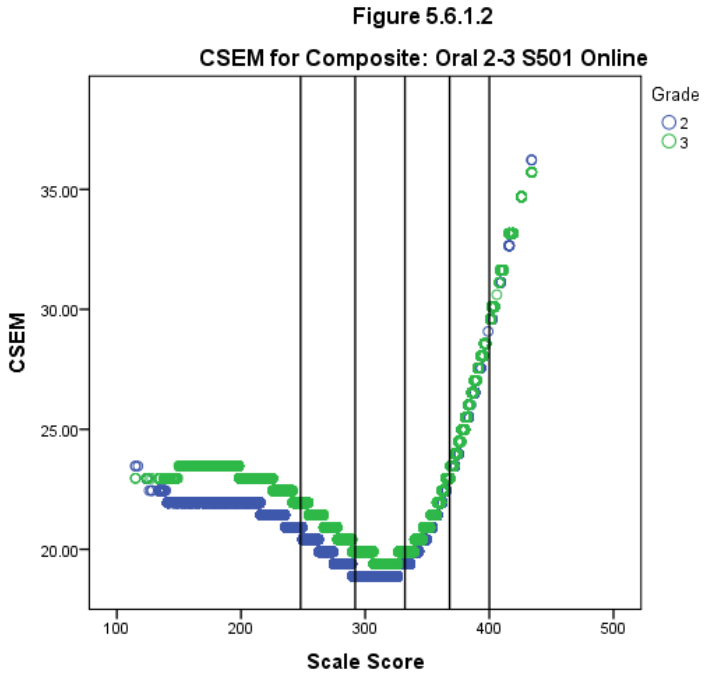
5.6.1.1 *Grade 1*

Figure 5.6.1.1



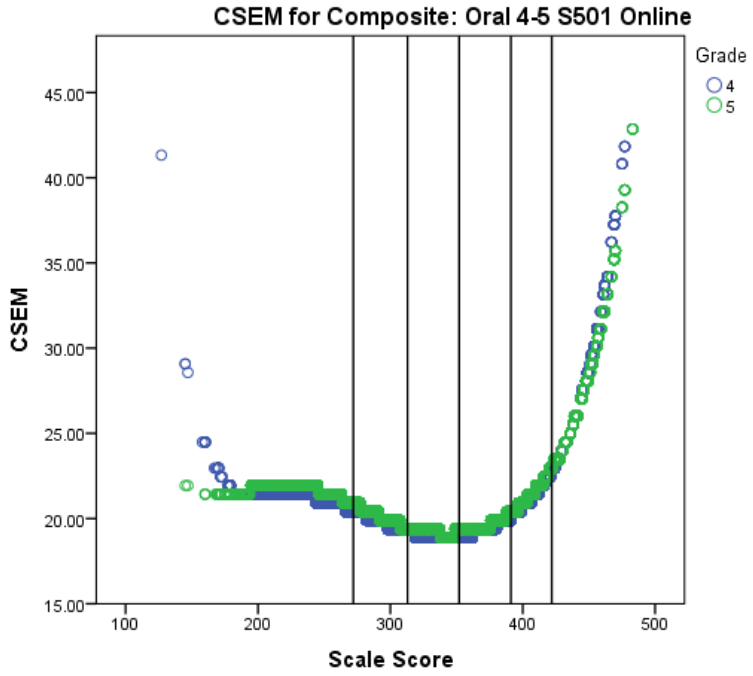
5.6.1.2 *Grades 2–3*

Figure 5.6.1.2



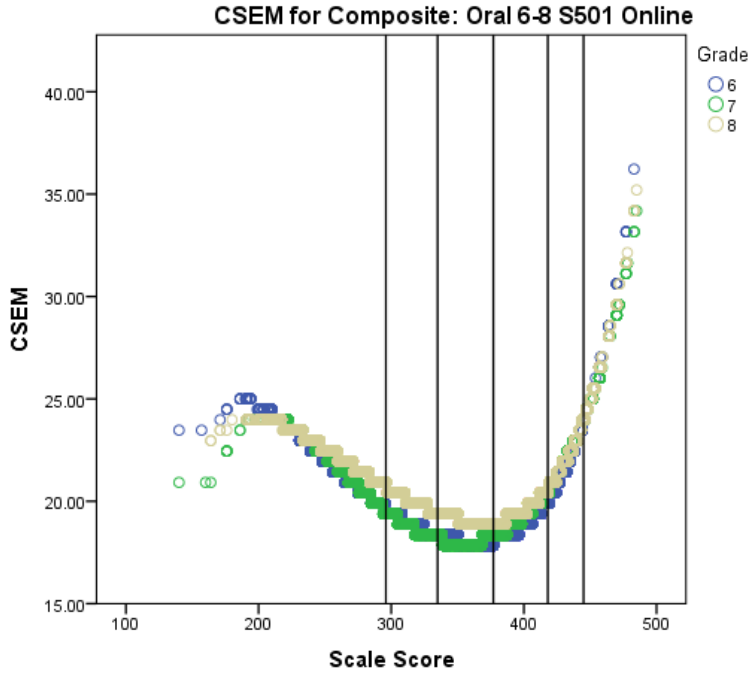
5.6.1.3 *Grades 4–5*

Figure 5.6.1.3



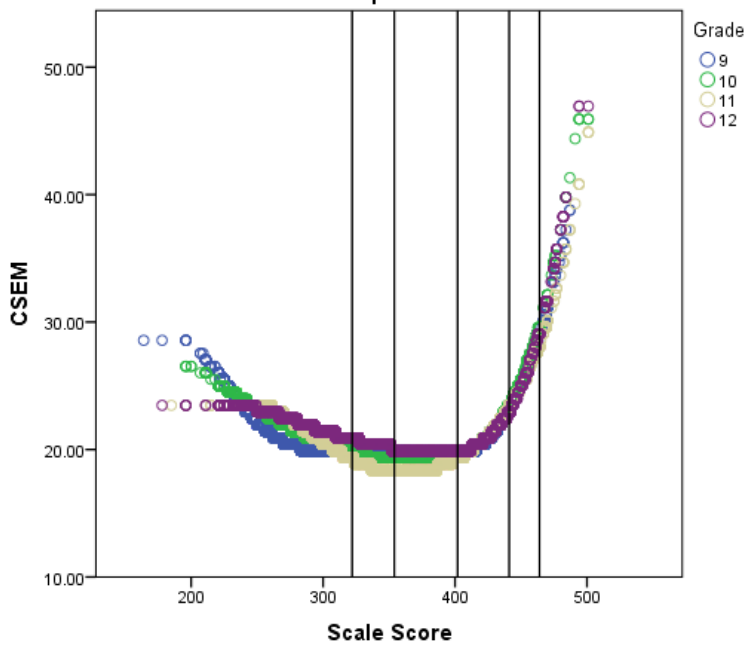
5.6.1.4 *Grades 6–8*

Figure 5.6.1.4



5.6.1.5 *Grades 9–12*

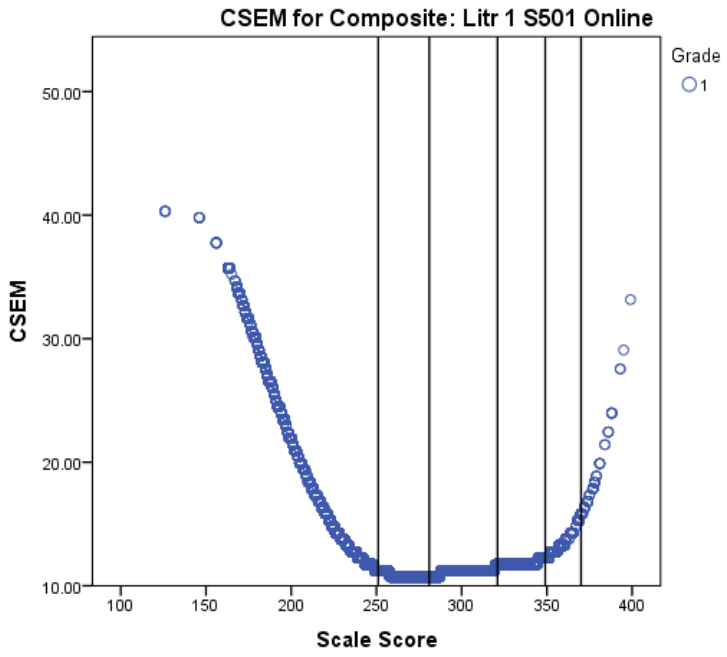
Figure 5.6.1.5



5.6.2 **Literacy**

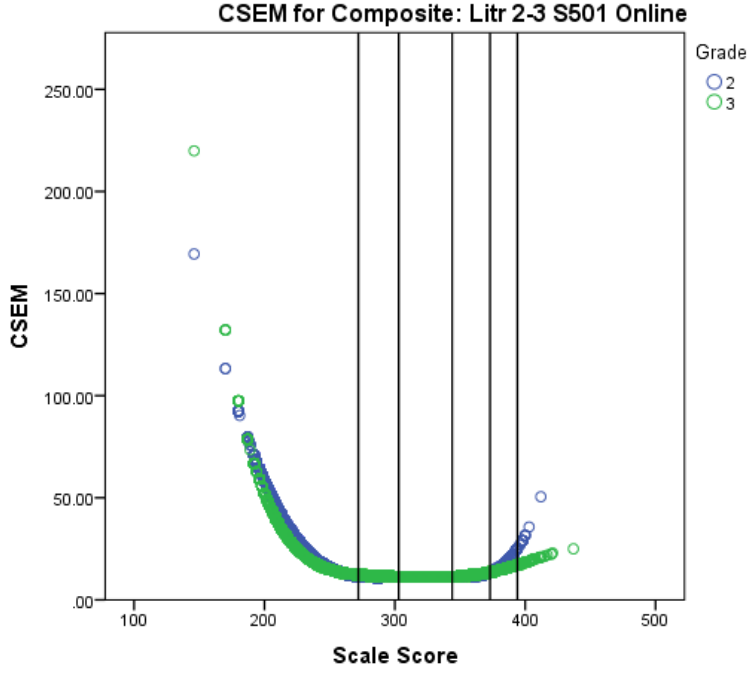
5.6.2.1 *Grade 1*

Figure 5.6.2.1



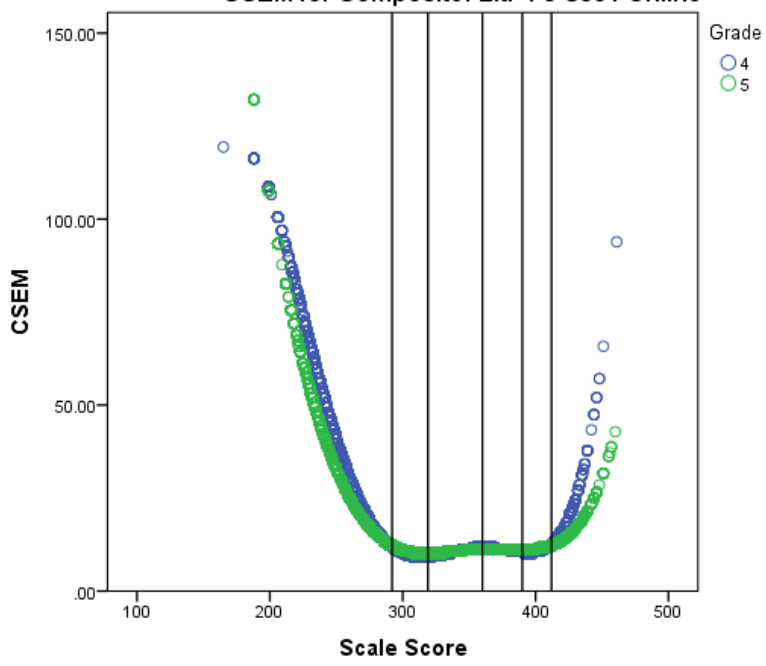
5.6.2.2 *Grades 2–3*

Figure 5.6.2.2



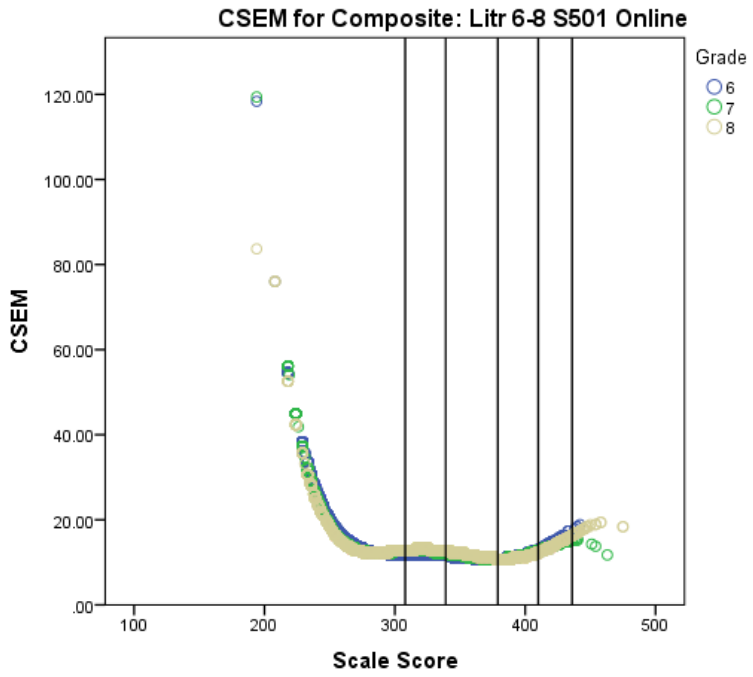
5.6.2.3 *Grades 4–5*

Figure 5.6.2.3



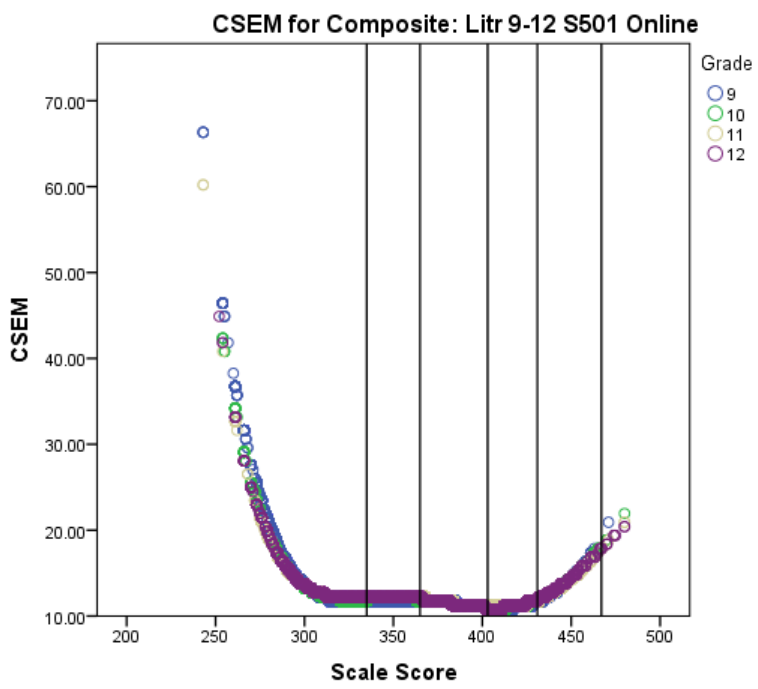
5.6.2.4 *Grades 6–8*

Figure 5.6.2.4



5.6.2.5 *Grades 9–12*

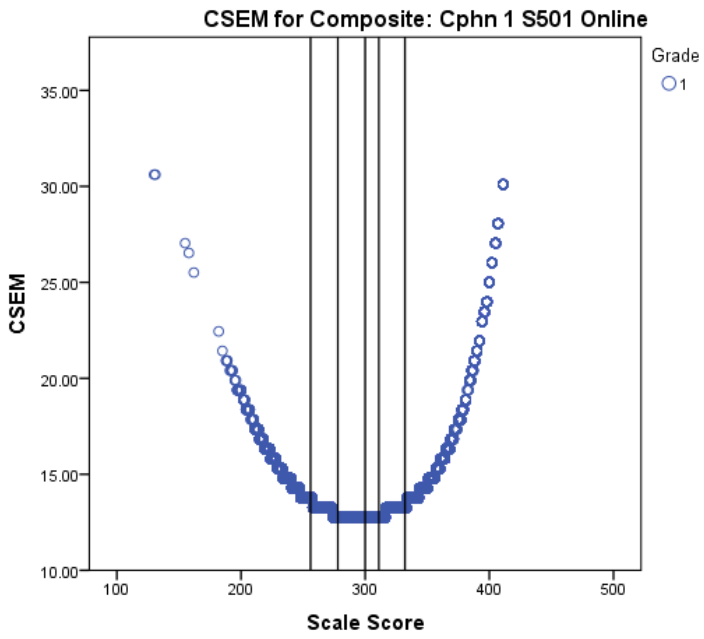
Figure 5.6.2.5



5.6.3 **Comprehension**

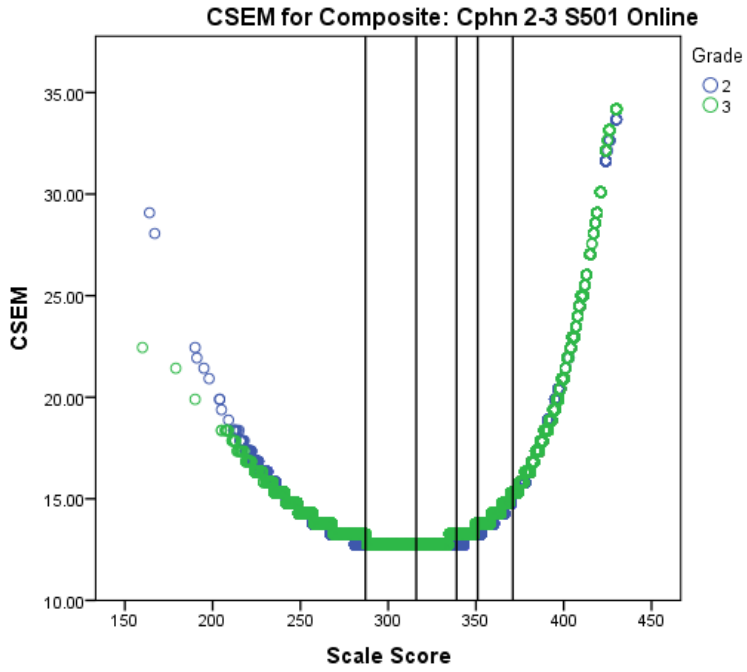
5.6.3.1 *Grade 1*

Figure 5.6.3.1



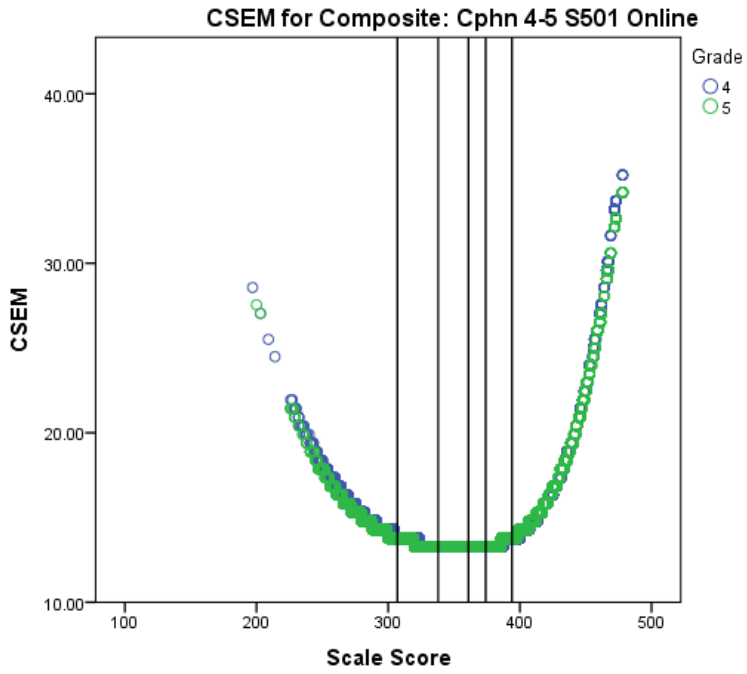
5.6.3.2 *Grades 2–3*

Figure 5.6.3.2



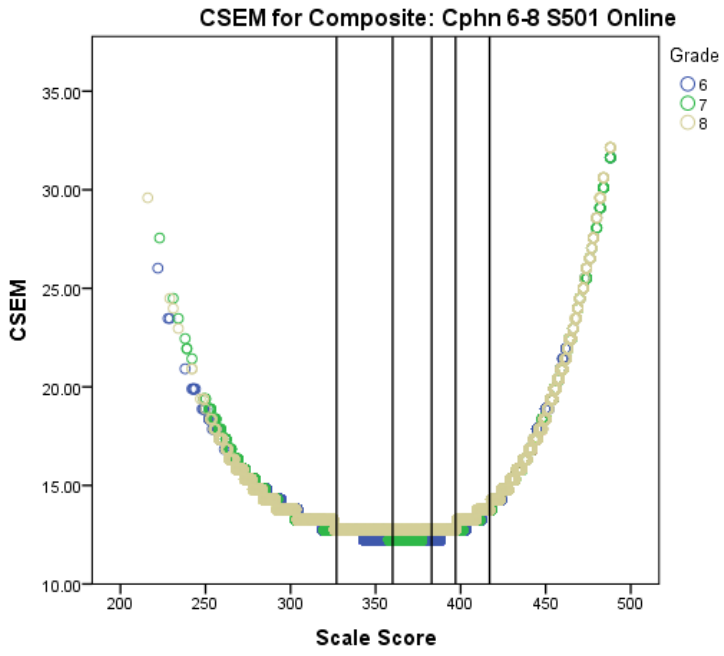
5.6.3.3 *Grades 4–5*

Figure 5.6.3.3



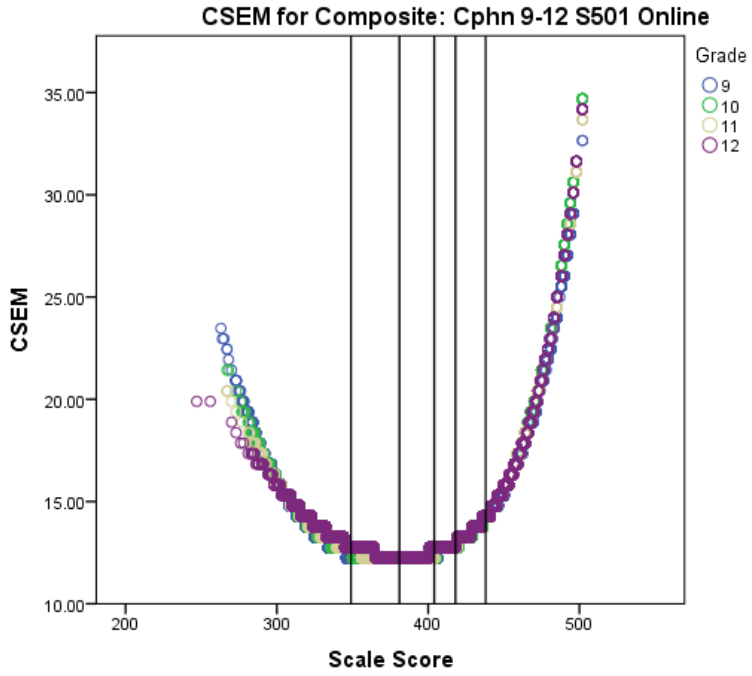
5.6.3.4 *Grades 6–8*

Figure 5.6.3.4



5.6.3.5 *Grades 9–12*

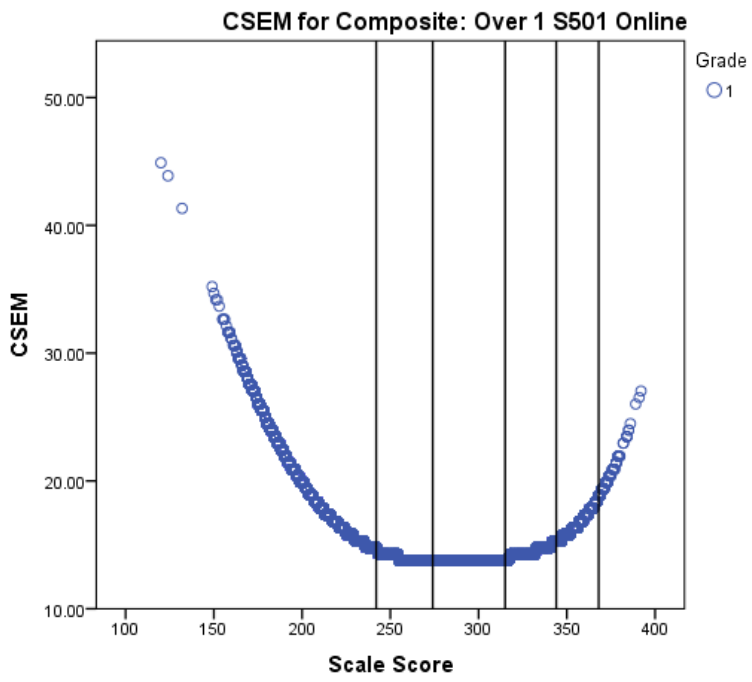
Figure 5.6.3.5



5.6.4 **Overall**

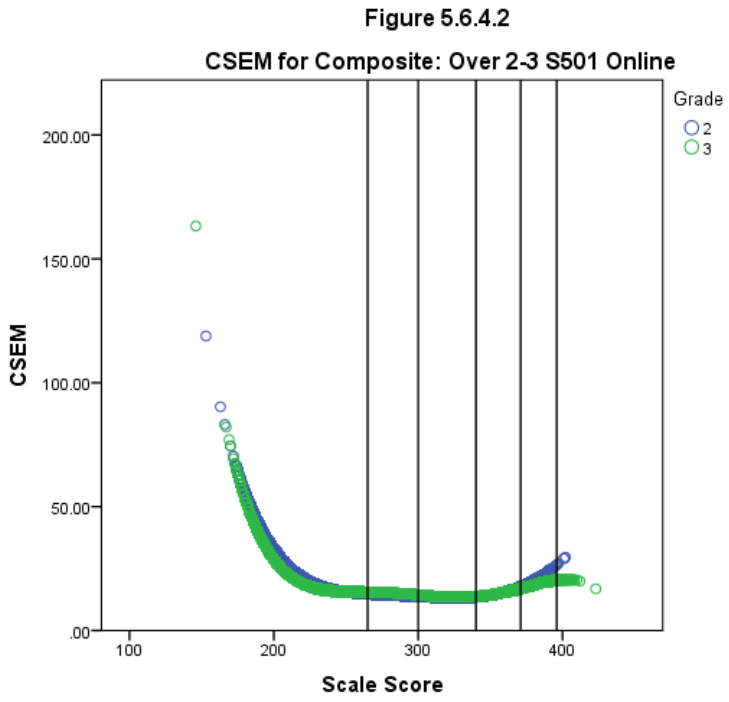
5.6.4.1 *Grade 1*

Figure 5.6.4.1



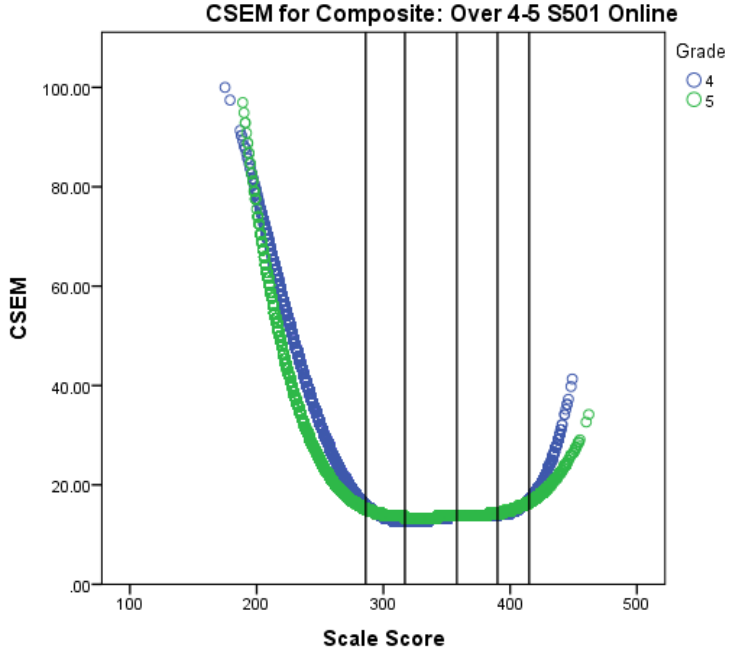
5.6.4.2 *Grades 2–3*

Figure 5.6.4.2



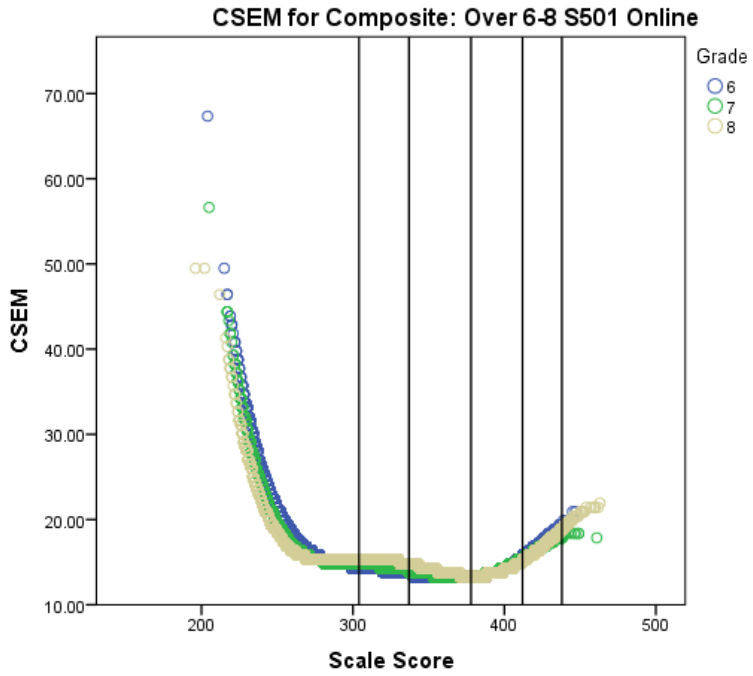
5.6.4.3 *Grades 4–5*

Figure 5.6.4.3



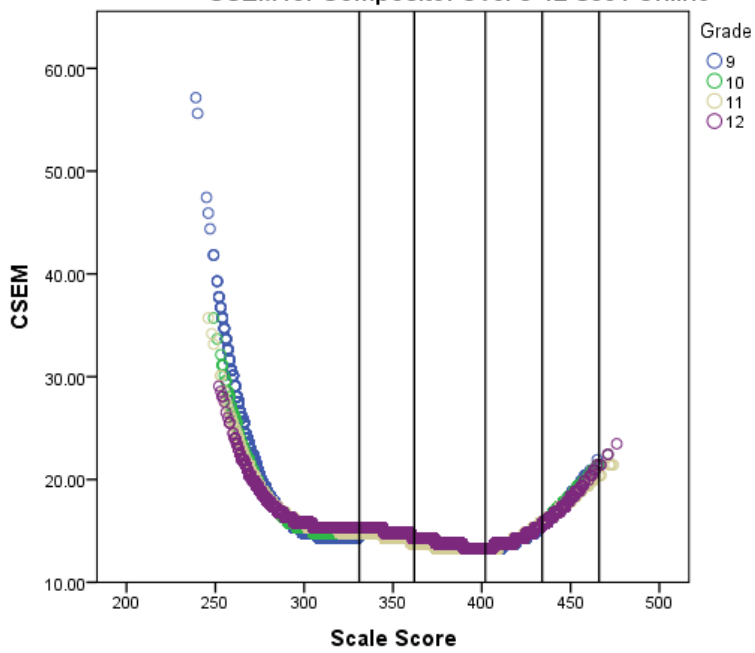
5.6.4.4 *Grades 6–8*

Figure 5.6.4.4



5.6.4.5 *Grades 9–12*

Figure 5.6.4.5



5.7 Accuracy and Consistency of Composites

One of the main purposes of the WIDA ACCESS program is to identify the English language proficiency level of students with respect to the WIDA ELD Standards. Because of the emphasis on the classification of student performance, a psychometric property of interest is how accurately and consistently ACCESS composite 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 incorporate one or more of the domains and composite scores in making accountability decisions, all WIDA Consortium states use the **Overall composite** as the primary score in making classification decisions about students. Therefore, it is especially important to examine the accuracy and consistency of the classifications based on the Overall composite 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 outlined by Livingston and Lewis (1995) and Young and Yoon (1998), 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 classification accuracy and consistency of the domains versus composites is that in order to estimate classification accuracy and consistency of the composites, we first estimated the reliability of the composite scores using a stratified Cronbach's alpha coefficient, as described in Section 5.4.

For each test domain, we present three tables. The first provides the overall accuracy and the overall consistency for each grade level. The second provides the classification accuracy at the cut score for each grade level. The third provides the classification consistency at the cut score for each grade level.

If the overall and marginal classification accuracy and consistency indices cannot be estimated because there are fewer than 200 students in the proficiency level, we collapsed the affected proficiency level category with the category below it and placed 'N/A' in the table for the affected proficiency level.

As noted in Section 5.4, there has been very little guidance for the ideal or expected levels of decision consistency and accuracy needed for educational assessments. We provide detail on the range of these statistics, by each composite, highlighting the grade level with the lowest classification accuracy and consistency of the composites for test users and policy makers. Since overall accuracy and consistency statistics are a summary of the degree of classification accuracy and consistency across all proficiency level cut points, the marginal classification accuracy and consistency for these grades were further examined to identify the specific source(s) of low classification accuracy and consistency.

For the Oral composite, as shown in Table 5.7.1.1, overall classification accuracy ranged from 0.629 to 0.749 and overall classification consistency ranged from 0.521 to 0.656 across grades. The lowest overall classification accuracy and consistency values were found for students in Grade 5.

For the Literacy composite, as shown in Table 5.7.2.1., overall classification accuracy ranged from 0.669 to 0.756 and overall classification consistency ranged from 0.555 to 0.661 across grades. The lowest overall classification accuracy and consistency values were found for students in Grade 5.

For the Comprehension composite, as shown in Table 5.7.3.1, overall classification accuracy ranged from 0.642 to 0.716 and overall classification consistency ranged from 0.535 to 0.629 across grades. The lowest overall classification accuracy and consistency values were found for students in Grade 1.

For the Overall composite, as shown in Table 5.7.4.1, overall classification accuracy ranged from 0.719 to 0.811 and overall classification consistency ranged from 0.627 to 0.737 across grades. The lowest overall classification accuracy and consistency values were found for students in Grade 5.

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

From an accountability perspective, the most important information for test users and policy makers to examine is the marginal classification accuracy and consistency. We summarize the range of the marginal classification accuracy and consistency of composites across grades and highlight the grade level with the lowest marginal classification accuracy and the lowest consistency, by composite.

For the Oral composite, classification accuracy indices at the cut ranged from 0.827 to 0.997 (Table 5.7.1.2) and classification consistency at the cut ranged from 0.761 to 0.997 (Table 5.7.1.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 5 at the PL 4/PL 5 cut. Note that Grade 5 was also identified as having the lowest overall classification accuracy and consistency in the Oral composite. The low marginal classification accuracy and consistency at the PL 4/PL 5 cut appeared to have contributed to its low overall classification accuracy and consistency. However, it should be noted that the marginal classification accuracy and consistency for Grade 5 Oral composite are in the high 70's and low 80's.

For the Literacy composite, classification accuracy indices at the cut ranged from 0.860 to 0.999 (Table 5.7.2.2) and classification consistency at the cut ranged from 0.803 to 0.999 (Table 5.7.2.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 5 at the PL 3/PL 4 cut. Note that Grade 5 was also identified as having the lowest overall classification accuracy and consistency in the Literacy composite. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut 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 for Grade 5 Literacy composite are still in the 80's.

For the Comprehension composite, classification accuracy indices at the cut ranged from 0.901 to 0.987 (Table 5.7.3.2) and classification consistency at the cut ranged from 0.864 to 0.982 (Table 5.7.3.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 1 at the PL 3/PL 4 cut. Note that Grade 1 was also identified as having the lowest overall classification accuracy and consistency in the Comprehension composite. The low marginal classification accuracy and consistency at the PL 3/PL 4 cut 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 for Grade 1 Comprehension are still in the high 80's and low 90's.

For the Overall composite, classification accuracy indices at the cut ranged from 0.859 to 0.987 (Table 5.7.4.2) and classification consistency at the cut ranged from 0.828 to 0.986 (Table 5.7.4.3). The lowest marginal classification accuracy and consistency values were found for students in Grade 5 at the PL 4/PL 5 cut. Note that Grade 5 was also identified as having the lowest overall classification accuracy and consistency in the Overall composite. The low marginal classification accuracy and consistency at the PL 4/PL 5 cut 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 for Grade 5 Overall composite are still in the 80's.

The results from the overall and marginal classification accuracy and consistency statistics provided similar findings. Grade 5 had the lowest overall and marginal classification accuracy and consistency in three of four composites (Oral, Literacy, and Overall), and Grade 1 had the lowest overall and marginal classification accuracy and consistency in the Comprehension composite. In addition, the lowest marginal classification accuracy and consistency of the composites occurred at the PL 3/PL 4 and PL 4/PL 5 cut points. This finding is consistent with previous research (Lee et al., 2000), in that classification accuracy and consistency at cut points in the middle of the proficiency level range are lower than those in the lower and upper ends. A higher number of proficiency levels typically results in cut scores that are closer to each other than if a smaller number of proficiency levels were used. Classification accuracy and consistency are expected to vary for different ability levels due to variation in measurement accuracy. The further away the scores are from the cut scores, the smaller the classification errors would be or the more accurate the classification decisions would be. With a large number of proficiency levels, there are more students near the cut scores 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 PLs 3 and 4 occupy relatively narrow ranges on the ability scale compared with other proficiency levels, the classification accuracy and consistency for the 3/4 and 4/5 cuts are lower than for other cuts.

There has been very little guidance for the ideal or expected levels of decision consistency and accuracy needed for educational assessments that use composite scores. From an accountability

perspective, the most important information for test users and policy makers to examine is the marginal classification accuracy and consistency. The marginal classification accuracy and consistency indices were at or above 0.800 for all composites except for the Oral composite. The lowest marginal classification consistency for the Oral composite was 0.761 for Grade 5. Additionally, the marginal classification accuracy and consistency indices were at or above 0.828 for the Overall composite, where the major accountability decisions are being made.

5.7.1 **Oral**

Table 5.7.1.1

Overall Accuracy and Consistency of Classification Indices: Oral S501 Online

Grade	Accuracy	Consistency
1	0.670	0.560
2	0.680	0.573
3	0.673	0.564
4	0.648	0.543
5	0.629	0.521
6	0.731	0.632
7	0.709	0.605
8	0.698	0.591
9	0.742	0.645
10	0.742	0.645
11	0.744	0.649
12	0.749	0.656

Table 5.7.1.2

Classification Accuracy Indices at Cut Score Level: Oral S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.966	0.933	0.887	0.901	0.979
2	0.964	0.922	0.887	0.910	0.993
3	0.971	0.935	0.881	0.885	0.994
4	0.988	0.972	0.923	0.859	0.904
5	0.984	0.966	0.919	0.827	0.928
6	0.981	0.948	0.891	0.919	0.990
7	0.971	0.932	0.889	0.927	0.988
8	0.964	0.928	0.890	0.924	0.987
9	0.939	0.911	0.917	0.975	0.997
10	0.948	0.913	0.907	0.974	0.997
11	0.949	0.912	0.908	0.974	N/A
12	0.945	0.910	0.912	0.981	N/A

Table 5.7.1.3

Classification Consistency Indices at Cut Score Level: Oral S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.953	0.903	0.843	0.860	0.973
2	0.949	0.889	0.841	0.880	0.992
3	0.957	0.908	0.831	0.849	0.994
4	0.983	0.959	0.890	0.804	0.880
5	0.978	0.950	0.887	0.761	0.906
6	0.973	0.924	0.848	0.889	0.990
7	0.958	0.902	0.845	0.899	0.987
8	0.949	0.897	0.846	0.895	0.985
9	0.914	0.875	0.882	0.967	0.997
10	0.926	0.876	0.869	0.966	0.997
11	0.927	0.876	0.869	0.968	N/A
12	0.922	0.873	0.875	0.977	N/A

5.7.2 **Literacy**

Table 5.7.2.1

Overall Accuracy and Consistency of Classification Indices: Litr S501 Online

Grade	Accuracy	Consistency
1	0.754	0.659
2	0.739	0.640
3	0.724	0.620
4	0.683	0.571
5	0.669	0.555
6	0.756	0.661
7	0.739	0.639
8	0.720	0.615
9	0.723	0.621
10	0.733	0.632
11	0.739	0.638
12	0.744	0.644

Table 5.7.2.2

Classification Accuracy Indices at Cut Score Level: Litr S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.908	0.886	0.967	0.992	N/A
2	0.959	0.904	0.892	0.983	N/A
3	0.969	0.927	0.861	0.966	0.999
4	0.976	0.947	0.860	0.898	0.983
5	0.972	0.942	0.860	0.893	0.985
6	0.947	0.898	0.920	0.992	N/A
7	0.939	0.894	0.919	0.986	N/A
8	0.930	0.896	0.909	0.983	N/A
9	0.938	0.906	0.911	0.968	N/A
10	0.952	0.909	0.905	0.967	N/A
11	0.954	0.906	0.907	0.972	N/A
12	0.950	0.899	0.916	0.980	N/A

Table 5.7.2.3

Classification Consistency Indices at Cut Score Level: Litr S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.868	0.840	0.955	0.991	N/A
2	0.940	0.866	0.848	0.978	N/A
3	0.955	0.896	0.805	0.952	0.999
4	0.964	0.923	0.804	0.863	0.981
5	0.960	0.917	0.803	0.855	0.983
6	0.925	0.856	0.885	0.990	N/A
7	0.915	0.851	0.885	0.981	N/A
8	0.902	0.853	0.873	0.973	N/A
9	0.914	0.866	0.874	0.955	N/A
10	0.933	0.871	0.867	0.953	N/A
11	0.936	0.868	0.869	0.960	N/A
12	0.928	0.858	0.881	0.972	N/A

5.7.3 **Comprehension**

Table 5.7.3.1

Overall Accuracy and Consistency of Classification Indices: Cphn S501 Online

Grade	Accuracy	Consistency
1	0.642	0.535
2	0.673	0.569
3	0.664	0.564
4	0.716	0.629
5	0.691	0.601
6	0.697	0.596
7	0.689	0.589
8	0.683	0.586
9	0.707	0.611
10	0.699	0.601
11	0.698	0.600
12	0.701	0.602

Table 5.7.3.2

Classification Accuracy Indices at Cut Score Level: Cphn S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.943	0.910	0.901	0.916	0.953
2	0.965	0.924	0.912	0.917	0.945
3	0.961	0.930	0.914	0.911	0.931
4	0.987	0.964	0.935	0.914	0.904
5	0.979	0.958	0.926	0.906	0.906
6	0.967	0.935	0.910	0.920	0.959
7	0.959	0.929	0.911	0.925	0.956
8	0.952	0.930	0.915	0.925	0.951
9	0.946	0.927	0.930	0.938	0.958
10	0.954	0.929	0.926	0.930	0.951
11	0.957	0.929	0.924	0.926	0.951
12	0.953	0.927	0.925	0.930	0.959

Table 5.7.3.3

Classification Consistency Indices at Cut Score Level: Cphn S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.920	0.873	0.864	0.882	0.933
2	0.950	0.894	0.876	0.883	0.921
3	0.945	0.900	0.878	0.875	0.902
4	0.982	0.949	0.907	0.879	0.866
5	0.971	0.939	0.895	0.870	0.867
6	0.955	0.907	0.876	0.888	0.940
7	0.943	0.899	0.877	0.895	0.937
8	0.933	0.900	0.882	0.894	0.929
9	0.925	0.897	0.901	0.913	0.941
10	0.935	0.900	0.895	0.902	0.930
11	0.940	0.900	0.892	0.898	0.930
12	0.934	0.898	0.894	0.903	0.940

5.7.4 **Overall**

Table 5.7.4.1

Overall Accuracy and Consistency of Classification Indices: Over S501 Online

Grade	Accuracy	Consistency
1	0.805	0.727
2	0.793	0.713
3	0.781	0.699
4	0.726	0.637
5	0.719	0.627
6	0.811	0.737
7	0.794	0.713
8	0.780	0.693
9	0.796	0.716
10	0.802	0.724
11	0.806	0.730
12	0.809	0.734

Table 5.7.4.2

Classification Accuracy Indices at Cut Score Level: Over S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.955	0.909	0.955	0.987	N/A
2	0.972	0.932	0.912	0.978	N/A
3	0.978	0.948	0.891	0.965	N/A
4	0.986	0.968	0.914	0.873	0.980
5	0.983	0.964	0.912	0.859	0.985
6	0.974	0.935	0.915	0.987	N/A
7	0.965	0.928	0.919	0.982	N/A
8	0.957	0.924	0.921	0.978	N/A
9	0.954	0.931	0.932	0.980	N/A
10	0.962	0.932	0.929	0.979	N/A
11	0.964	0.932	0.930	0.980	N/A
12	0.961	0.926	0.935	0.987	N/A

Table 5.7.4.3

Classification Consistency Indices at Cut Score Level: Over S501 Online

Grade	PL 1/2	PL 2/3	PL 3/4	PL 4/5	PL 5/6
1	0.934	0.872	0.934	0.986	N/A
2	0.959	0.904	0.875	0.973	N/A
3	0.968	0.925	0.847	0.956	N/A
4	0.980	0.954	0.880	0.832	0.979
5	0.976	0.948	0.876	0.828	0.984
6	0.963	0.908	0.880	0.986	N/A
7	0.950	0.899	0.886	0.978	N/A
8	0.940	0.893	0.889	0.969	N/A
9	0.937	0.901	0.904	0.973	N/A
10	0.947	0.904	0.899	0.974	N/A
11	0.950	0.903	0.900	0.976	N/A
12	0.944	0.896	0.907	0.986	N/A

6 Quality Control

6.1. Content Development Quality Control

The Center for Applied Linguistics (CAL) utilizes educators and other consultants at a number of phases throughout the test development cycle. These educators and consultants are recruited, vetted, and trained by CAL and/or WIDA and make crucial contributions to these phases of the test development cycle. The phases of development in which educators or consultants are involved, as well as the procedures and criteria for recruitment and training, are described below.

Theme Generation

During theme generation, CAL and WIDA recruit educators to generate raw ideas to be used in new item development. Educators with ESL or content-area expertise and two or more years of teaching experience in a WIDA state (in the grade cluster for which they will generate themes) are invited to participate. Recruitment also focuses on a geographical distribution of educators from across the consortium. Upon selection, educators participate in a short training that introduces the theme generation process, along with how to understand the item specifications that they use to generate themes.

Item Writing

CAL recruits professional item writers to generate raw item/task content based on the ideas from theme generation. To recruit item writers, CAL has a standing announcement on its website asking prospective item writers to submit their 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 & Sensitivity and Content Review. These are technically two separate reviews, although a single recruitment effort is conducted by WIDA, and the reviews occur consecutively in a single week (generally 3 days for Content review followed by 2 days for Bias & Sensitivity review). As with other reviews, educators for Content review must have at least 2 years of ESL teaching experience (with a preference for content-area experience as well). Recruitment also focuses on selecting educators with a variety of cultural and linguistic backgrounds and obtaining a geographical distribution of educators from across the consortium. Recruitment for Bias & Sensitivity review focuses on selecting educators with culturally and linguistically diverse backgrounds who have experience interacting with English learners from a range of cultural, regional, religious, linguistic, ethnic, and socioeconomic backgrounds.

At the beginning of both Bias & Sensitivity and Content review meetings, CAL and WIDA staff conduct an intensive training to orient the reviewers to the specific purpose of the review (Bias & Sensitivity or Content), how to use the review checklist and what to look for in the review, and the procedures and security protocols for the review. Then, the reviews are conducted in breakout groups by grade cluster (or combinations of grade clusters; for example, Bias & Sensitivity review of Grade 1 and Grades 2–3 is often combined). Although Bias & Sensitivity and Content reviews are generally held in -person, the reviews for the Writing domain occur virtually each year due to timeline constraints. For both the in-person and virtual contexts, CAL and WIDA facilitators are present in each breakout group to guide the educators in their reviews of the materials.

Writing Tryouts

All tasks in the Writing domain are subject to tryouts in the field. The Writing tryouts only occur once the tasks have been through a thorough Bias & Sensitivity and Content review and subsequent revision. CAL and WIDA recruit educators who are willing to administer the Writing tasks to their students; these educators are classroom ESL or content teachers who work with ELs. All students who participate are required to have parent/guardian consent.

Once the students complete the Writing tasks, both the students and educators fill out questionnaires. Student questionnaires focus on whether the students understood the task, their engagement with the task, and their ability to complete the task; educator surveys ask the teachers to evaluate the effectiveness of the task input, the appropriateness of the task, the comparability of the task with other classroom-based writing tasks, and the ability of the students to complete the task.

CAL provides the teachers with a number of documents outlining the procedures for administering the tasks, recording student responses to the tasks, recording student and teacher responses to the questionnaires, and protecting the personally identifiable information of the students. CAL staff are also available throughout the tryouts process to answer any questions the teachers might have. Following the Writing tryouts, CAL specialists review the writing responses both qualitatively and quantitatively, providing WIDA with a report on how the Writing tasks performed.

6.2. Test Administration Quality Control

This section describes how WIDA monitors test administration to ensure standardized test administration procedures are implemented with fidelity across districts and schools. To support standardized administrations, WIDA provides test administrators with a series of resources, such as a Test Administration Manual, a training course, and a Test Administration script for each assessment.

Qualifications of Test Administrators

Before, during, and after a state's testing window, educators hold various roles to ensure all tasks are carried out for successful test administration. These roles include Test Coordinators at the district and school level and Test Administrators. The Test Administrator administers and monitors the test, and is also responsible for managing student data prior to, during, and after testing.

WIDA has worked directly with each state education agency to develop the ACCESS for ELLs Checklist for the school year. This list highlights all tasks that need to be completed before, during, and after testing within a school or district and outlines which tasks are assigned to Test Coordinators at the district and school level and to Test Administrators. It also provides additional guidance that a state expects test administrators to follow as they prepare for and administer the ACCESS for ELLs suite of assessments.

Test Administrators are responsible for reviewing each state's checklist in detail prior to completing any training and for working with the district or school Test Coordinator to complete these tasks. The state's checklist can be found in the training course and on each state's WIDA webpage at www.wida.us/membership/states.

The training course within the WIDA Secure Portal (https://www.wida.us/login.aspx) is where educators can access both training to become certified to administer ACCESS for ELLs as well as additional materials and resources to assist administrators and coordinators before, during, and after each state's testing window. WIDA user accounts provide access to the training course and Facilitator Toolkit within the WIDA Secure Portal. Educators must pass an administration quiz at the end of the training with a score of 80% or higher. WIDA recommends taking the quiz immediately after completing the training. There is no limit to the number of times educators can attempt the quiz. Once individuals pass an administration quiz, training certificates within the WIDA Secure Portal are updated to reflect their status as a certified Test Administrator for that component of the assessment suite.

Paper Testing (for Writing Grades 1–3)

Depending on state, district, and school policy, not all Test Administrators will be responsible for initially labeling and/or bubbling booklets. However, it is the responsibility of all Test Administrators and Test Coordinators to ensure that correct and complete information is either labeled or bubbled in each student booklet. Each state's ACCESS for ELLs Checklist has more information on who is responsible for each task related to materials management in the state.

To ensure all booklets have the detailed and necessary information needed to score, all Test Administrators must adhere to the following:

- Prior to administration
    - Review labels and/or bubbled information to ensure all student information is accurate.
    - Complete labeling or bubbling if needed.
- During administration
    - Distribute the test booklets, as applicable, to the correct students.
    - Verify that students have been given their assigned booklet.
- Immediately following administration
    - Collect all material from all students.
    - Review student test booklets once more for any errors or discrepancies in student information.
    - Confirm all necessary fields are completed and all necessary labels are correctly adhered to student test booklets.
    - Ensure all booklets are in proper condition to be returned, with no loose or damaged pages.

- Return test materials to a Test Coordinator, or store the booklets in a secure area until they can be handed over to a Test Coordinator.

Failure to address incorrect, missing, or incomplete booklet information and labels may result in late reporting or no student score. In addition, the WIDA Consortium's national research agenda relies on complete and accurate student demographic data to inform the field and benefit English language learners.

When preparing test materials for return to DRC, test administrators need to confirm that any booklet that contains student response information has either a Pre-ID Label or a District/School Label with bubbled student information. If a booklet is unused, there is no need to place any labels on the booklet. Placing a label on a booklet will cause it to be processed (and either scored, if the label is a Pre-ID or School/District label, or not scored, if it is a Do Not Process label).

6.3. Rater Quality Control

Rater Training

Students who take the ACCESS for ELLs Paper Speaking test have their spoken responses scored by the Test Administrator who administered the Speaking test. Another term for this Test Administrator is *rater*. Raters must be trained and certified, so we can be confident that they interpret students' spoken language consistently and fairly and that the scores are reported according to the WIDA English language proficiency standards. WIDA provides several different types of resources to support raters' training and reliability.

Students who take ACCESS for ELLs Online have their spoken responses digitally recorded and then scored centrally by DRC's trained raters. Students who take ACCESS for ELLs Paper have their spoken responses scored in real time by the Test Administrator who administers the Speaking test. In both cases, it is important that the individual who scores the spoken responses is trained and certified.

WIDA provides a series of training modules in the Secure Portal on the WIDA website. ACCESS for ELLs Speaking test raters should complete three core modules:

1. Overview and Test Structure
2. Speaking Assessment Scoring Practice
3. Speaking Assessment Recommended Practice

WIDA strongly recommends that all new raters complete all three of these modules. These modules provide a comprehensive introduction to the ACCESS for ELLs Speaking test and the opportunity to learn how to score students' spoken English reliably using the ACCESS for ELLs 2.0 Speaking Scoring Scale.

In addition to the modules described above, WIDA also releases supplemental training materials each year to refamiliarize experienced raters with the Speaking Scoring Scale and introduce new Speaking tasks and sample responses for the coming year. These materials, called Supplemental Training for the Speaking Assessment, reflect the Speaking tasks that will appear on the test in the current year. WIDA recommends that all raters (new and experienced) engage with these supplementary materials at the start of each scoring season. Reading and reviewing these materials will help raters maintain their reliability from year to year and contribute to the fairness of test scores awarded to all students.

Rater Certification

After completing the training modules described in the section above, new raters should take the relevant certification quiz. WIDA provides two quizzes: one for raters who will evaluate students in Grades 1–5 and another for raters who will evaluate students in Grades 6–12. Raters should take the appropriate quiz.

The purpose of the quiz is to ensure that raters have internalized the Speaking Scoring Scale and can apply it consistently. Only raters who pass the quiz(zes) should administer and score the ACCESS for ELLs 2.0 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 2.0 Speaking test
- ✓ Raters do not evaluate their own students, if at all possible
- ✓ Rater reliability and/or score point distributions are monitored regularly

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 quality control of all score reports; a secondary district, for quality control of State Frequency Reports only; and a tertiary district for quality control of paper-based tests only.

After an introductory presentation, which includes details of the quality control processes undertaken by DRC and WIDA and instructions on using the data entry tools, panelists begin by confirming the scoring of ACCESS Paper. Using the information in the State Student Response file, panelists enter the grade level, grade level cluster, tier, the Listening and Reading responses, and the Speaking and Writing scores into the data entry tool. The tool then calculates the student's raw scores and, using a series of look-ups, the student's scale score, proficiency level score, and confidence bands for all domains and composites. Panelists check student scores on the Individual Student Reports against those calculations. Any discrepancies are brought to the attention of the WIDA facilitator who investigates and, if there seems to be an issue with the report (rather than the data entry or data entry tool), discusses the issue further with DRC.

The panelists follow a similar process with the Kindergarten ACCESS tests, but with the raw scores for these tests copied directly from the response booklets.

After checking the paper-based tests, panelists turn their attention to the score reports. Panelists first check both the demographic information and the student scores in the Individual Student Reports against the information in the Student Roster Reports. Again, any discrepancies are brought to the attention of the facilitator, who investigates and discusses the issue with DRC if necessary. Panelists use the verified Individual Student Reports to check the Student Roster Report. Once the Student Roster Report is verified, panelists use it to check the State Frequency Report; they then use the verified State Frequency Reports to check the District Frequency Report. Finally, panelists check the State Frequency Reports against verified District Frequency Reports from the primary district along with District Frequency Reports from the secondary district.

6.5. Data Forensic Quality Control

WIDA hired Caveon to perform data forensic analysis during the 2019–2020 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.

[3] Due to the COVID-19 pandemic, the 2020 Score Reporting quality control was conducted online, with only WIDA and DRC employees participating.

Data forensics analysis was performed after the administration window for the following administrations:

- December 2019 through Spring 2020 online multistage adaptive test administrations, Listening and Reading domains
- December 2019 through Spring 2020 paper fixed-form administrations, Listening and Reading domains

The analysis utilized several of Caveon's security statistics to detect evidence of whether the assessment instrument has been compromised through disclosure of the content. This analysis attempted to understand where and when disclosure of the test content may have occurred and what items and forms may have been affected. Results of this analysis might enable WIDA to take specific actions to limit the impact of disclosed content. Such actions may include

- Republishing or reworking items or forms
- Rotating disclosed items to limit their exposure
- Designing a republication or rotation strategy for future items and forms

Caveon security statistics were computed for each individual test instance. These data were aggregated or summarized at the group level. The aggregated statistics were compared against the population model.

Analysis of Tests

Caveon aggregated the data according to individual test forms using the security statistics to determine whether rates of detections by the security statistics were higher for certain test forms. For fixed-form paper tests, two forms—A and B/C—were analyzed. For the multistage adaptive test, there is a finite number of ways a student could progress through the test. Caveon analyzed each pathway as a separate form. Higher rates of security detections for a specific form of the test suggest that compromise of the form may have occurred.

Analysis of Items

Item security: In this portion of the analysis, the security of the items was evaluated using aberrance statistics. Aberrance statistics detect test-taking behaviors such as answering difficult items correctly but answering easy items incorrectly, or unusual patterns in the time taken to answer test items. In the absence of security issues, aberrant test taking is expected to be the result of poor or uneven test preparation, illness or other physical malady, mental and emotional distractions, and so forth. These factors usually result in lower levels of test performance. When aberrance is associated with higher performance, however, test fraud may have occurred, such as 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 item response theory model and adjusted for varying test-taker performance levels. This means that detected performance shifts are invariant to fluctuations in the test-taker population. When performance shifts indicate the item has become significantly easier, the item may have been disclosed. Items with significant performance shifts become candidates for revision or replacement. Item performance shifts were detected with a granularity of 1 week, where Monday to Sunday represents 1 week.

Analysis of Groups

Analysis by week: This analysis aggregates the data according to the week in which the test was taken to identify whether security threats and pass rates appeared to be more prevalent at certain times during the testing window. Increases in scores or security detections during certain periods of time suggest the content may have been disclosed at some point prior to that time. This analysis also includes a form-date grouping to determine if increasing security threats are associated with a particular form of the test. This analysis is performed for online and paper tests, where relevant test date data are provided.

Analysis of WIDA jurisdictions: Caveon analyzed WIDA member jurisdictions (states and districts) to determine whether rates of detections by the security statistics were higher for certain jurisdictions. This analysis is intended to detect whether compromise at the state or member jurisdiction level potentially occurred. This analysis is performed for online and paper tests.

Analysis of administration mode: Caveon aggregates the data according to administration mode (i.e., online versus paper) to determine if security threats are associated with the mode of testing.

Other Analyses

Analysis of mean score over time was used to identify whether mean scores increased over time during the testing window. Increases in scores over time suggest the content may have been disclosed during the testing window.

Findings of Data Forensic Analyses

Generally, no major data forensic anomalies were observed across WIDA states. There were some general findings and a few minor localized anomalies:

1. High rates of similar tests with associated score gains and a high rate of tests in large clusters suggest the presence of possible security violations in a district.
2. High rates of identical and/or perfect tests in two states suggest potential item compromise in these states.
3. For lower grades of the Reading Exam, examinees with better performance on old items than new items tended to have higher scores than those who did not exhibit a performance difference.

4. Paper-and-pencil exams had higher rates of identical and perfect tests than online exams. Within paper-and-pencil administrations, the Listening exam generally had higher rates of identical and perfect tests than the Reading exam.
5. Analysis of items suggested that some items may have been disclosed or become well known. This was especially prevalent among the online exams. However, if true, the disclosure appears to have occurred only among a low proportion of the examinees.
6. Analysis of test forms, test formats (i.e., administration mode), and test weeks did not find evidence of widespread item compromise or security violations. Mean scores were generally stable over the testing window.

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.

Bishop, K., Walker, C., Gocer Sahin, S., and Akanda, M. (2020). *DIF study by disability status in EL assessment*. WIDA Technical Report

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.

Elementary and Secondary Education Act of 1965, amended 2015. 20 USC §6301-8961.

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, NY: 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.

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17,* 221–240.

Kenyon, D. M. (2006). *Development and field test of ACCESS for ELLs®* (WIDA Consortium Technical Report No. 1). 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., Ho, P., Chapman, M., & Cook, H. G. (2020a). *Examination of reclassification decisions made for K–12 English learners: Survey report of Delaware* (WIDA Internal Report). Madison, WI: WIDA at the Wisconsin Center for Education Research.

Kim, A., Ho, P., Chapman, M., & Cook, H. G. (2020b). *Examination of reclassification decisions made for K–12 English learners: Survey report of Pennsylvania* (WIDA Internal Report). Madison, WI: WIDA at the Wisconsin Center for Education Research.

Kim, A., Tywoniw, R. L., & Chapman, M. (2020). *Performance of Technology-Enhanced Items in Grades 1-12 English Language Proficiency Assessments.*

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). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2002b, Autumn). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). 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. (n.d.). *Displacement measures*. Retrieved from http://www.winsteps.com/winman/displacement.htm

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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/

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*(4), 151–363.

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.

Rudner, L. (2001, Spring). Informed test component weighting. *Educational Measurement: Issues and Practice, 20*(1), 16–19.

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.

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, prekindergarten 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.

Wright, B.D. & Douglas, G.A. (1975). *Best test design and self-tailored testing*. Research memorandum, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* 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. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/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:

Keira Ballantyne, Ph.D.
Tanya Bitterman, M.A.
Caitlin Gdowski, M.A.
Yage (Leah) Guo, Ph.D.
Michele Kawood, M.S.Ed.
Justin Kelly, Ph.D.
Dorry M. Kenyon, Ph.D.
Nicholas Luzio Jr., B.S.
Erin Shaw-Meadow, M.Sc.
Samantha Musser, M.A.
Rachel Myers, M.S.
Yoon Ah Song, Ph.D.
Alice Tsai, M.S.
Shu Jing Yen, Ph.D.
Xin Yu, M.A.

From WIDA:

Mohammad Akanda, Ph.D.
Kyoungwon Bishop, Ph.D.
Sakine Göçer Sahin, Ph.D.