



# ELPA21 Item Development Process Report

Prepared by Maurice Cogan Hauck, Emilie Pooler,  
and David P. Anderson

Educational Testing Service

FINAL SUBMISSION

May 15, 2015

The contents of this document were developed under a grant from the US Department of Education. However, those contents do not necessarily represent the policy of the U.S. Department of Education and you should not assume endorsement by the Federal Government. To comply with Public Law 112-74, "Consolidated Appropriations Act, 2012", Division F – DEPARTMENTS OF LABOR, HEALTH AND HUMAN SERVICES, EDUCATION AND RELATED AGENCIES APPROPRIATIONS ACT, 2012, TITLE V – General Provisions, Sec. 505, December 23, 2011: 100% of the ELPA21 project are financed with Federal money; the Federal funds for ELPA21 total \$6,273,320; and 0% of this project is financed by non-governmental sources.

## Executive Summary

The ELPA21 Item Development Process Report summarizes the activities undertaken by Educational Testing Service (ETS) in 2014 on behalf of and in collaboration with the Council of Chief State School Officers (CCSSO) and the English Language Proficiency Assessment for the 21<sup>st</sup> Century (ELPA21) Consortium related to the design and development of a pool of test items for the ELPA21 assessment system.

The ELPA21 Consortium is a group of states organized to produce an assessment system that measures the language development of English language learners (ELLs). The system is intended to provide information that educational authorities in the consortium states can use to:

- Determine initial identification of ELLs (via the screener);
- Monitor ELLs' annual progress in the attainment of English for academic purposes;
- Measure districts' success in meeting accountability benchmarks per Title III of the No Child Left Behind Act; and
- Consideration for reclassifying students from ELL to Fluent English Proficient status.

The overall goal of the ELPA21 Consortium is to improve the measurement of ELL students' English proficiency by creating an assessment system based on an innovative set of English language proficiency (ELP) standards (CCSSO, 2014). Goals central to this item design and development effort included:

- Reflecting the values of the new ELP Standards, including a focus on the English needed for students to communicate and learn grade-appropriate content material in the academic contexts of English language arts, mathematics, and science;
- Taking advantage of contemporary approaches to computer-based assessment, including the use of a significant proportion of technology-enhanced (TE) test items;
- Foregrounding accessibility, ensuring that all test items are maximally accessible to all students, including students with disabilities;
- Supporting portability and interoperability, as the item pool must be amenable to handoff to other organizations for field test delivery and potentially other future use;
- Employing Evidence-Centered Design (ECD) to provide an intellectual underpinning that will serve as the basis for the assessment system's validity argument.

As the federal grant supporting the ELPA21 work had strict timelines, it was necessary for the ELPA21 item design and development work to be executed in a rapid and flexible manner, with a focus on ongoing problem solving. Following the organizational structure of the ELPA21 Consortium, the work was also conducted in a highly collaborative fashion; ETS collaborated consistently and openly with several ELPA21 Task Management Teams (TMTs), particularly those overseeing Item Acquisition and Development (IAD); Accessibility, Accommodations, and Administration (AAA); and Assessment Design and Scaling (ADS).

Essential contributions to the ELPA21 item pool were also made by educators from across the consortium states. These educators were positioned to make significant contributions because of their intimate knowledge of students' language skills and development, their insight into student interests and grade-appropriate topics, and their general ability to strengthen the link between assessment and instruction. Panels of educators played key roles in reviews of reading passages, in item writing, and in content and bias committee reviews of the item pool.

ECD looks on an educational assessment as “an evidentiary argument for reasoning what students say, do, or make in particular task situations as well as to generally claim what they can know, do, or have accomplished” (Mislevy, 2011, p. 6). ECD served as a framework for the process of conceptualizing, designing, and developing the ELPA21 item pool. ECD is commonly conceptualized as a series of five layers that constitute a progression from more abstract conceptualization to more specific and concrete instantiation: domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery.

A factor contributing to the necessity of a flexible, problem-solving approach to the ELPA21 item design and development work was that key products of the domain modeling layer (the claims, sub-claims, and Proficiency Level Descriptors [PLDs]) were being produced even as the schedule required work to be moving forward on the conceptual assessment framework and, at times, the assessment implementation. Although the necessity of working in multiple ECD levels simultaneously increased the complexity of the work, by the time of the delivery of the item pool, robust documentation supporting the domain modeling, conceptual assessment framework, and much of the assessment implementation layers of ECD had been produced, laying considerable groundwork for an effective ELPA21 validity argument.

While the conceptual work and the ECD-based documentation proceeded in an iterative fashion, the work of planning for production of and then developing items themselves was somewhat more linear. The process began with the development of assessment design documents (Test Blueprints, Reading Level Guidelines for passages, disaggregated PLDs needed to meet the needs of the assessment) and then task design documents (Item Specifications, draft rubrics for constructed-response [CR] tasks, draft scoring rules for TE items). Key decisions related to item banking and the metadata fields and values that would be captured for each item were also made at this stage, and an Editorial Style Guide and a Graphics Style Guide were produced, as well as initial versions of sample items.

As the work moved from the more conceptual stage to the production stage, initial tasks included the development of an overall item development plan as well as the acquisition of passages that would serve as the basis of test items in the reading domain. The item writing phase began with development of extensive item writer training materials (based directly on the initial (Stage 1) Item Specifications) and the recruitment and training of item writers, including a substantial cohort of educators from across the consortium states. The review process for all items was thorough and rigorous, including a series of internal reviews at ETS, reviews by representatives

of the TMTs, and large scale in-person reviews by content and bias review committees made up of educators from the consortium states. All reviews and comments received were then resolved in series of WebEx meetings with TMT members, at which point all content decisions related to the main item pool had been made.

At this point, the work turned to tasks associated with readying the main item pool for handoff to the field test delivery vendor (professional recording of needed audio files, embedding accessible content using the Accessible Portable Item Protocol [APIP], and executing quality control reviews for the items and their metadata), as well as generating Braille-ready items for potential paper delivery, including developing “twin” items for items in the main pool judged not to be accessible for students with visual impairments. The final task was a technical review of all items and their metadata and exporting to the field test delivery vendor.

The major deliverable produced at the end of the work described in this report was the design and development of the field test pool for the ELPA21 assessment system, sufficient to support initial field testing that will lead to the development of initial operational forms of a screener and summative assessment.

The pool delivered for field testing contained a total of 2,619 test items, 2,469 of which are intended for initial operational use, and 150 of which are experimental items (i.e., produced to be field tested but not intended for use in initial operational forms). Of the field-test-ready items, 1,178 or 45 percent contain accessible content authored via APIP. Of the items intended for initial operational use, 1,138 or 46 percent are TE items.

To support the accessibility goal, a supplementary pool of 415 “twin” items designed to enable administration to students with visual impairments was also developed.

In addition to the item pool itself, ETS also produced a range of ancillary deliverables, primarily documentation and tools that were necessary for the development of the item pool. These deliverables, which provide the basis for future ELPA21 item development efforts as well as contributing to the validity argument for ELPA21, include:

- Stage 1 Test Blueprints, providing information about anticipated test instances, item response types, and score points for the ELPA21 summative assessment;
- Reading Level Guidelines describing text complexity of ELPA21 reading passages;
- Draft disaggregated PLDs specific to each standard and domain;
- Stage 2 Item Specifications for each grade or grade band, providing detailed information about the design of all task types developed for ELPA21;
- Draft rubrics for CR task types;
- Draft scoring rules for TE task types, including draft partial credit scoring rules for those TE task types judged to be amenable to the awarding of partial credit;
- 193 sample items.

During the course of the item design and development work, ETS was also contracted to take on two additional pieces of work, resulting in the following deliverables:

- Design and execution of a cognitive laboratory study to assess how well students can interact with various item types, how well directions work, and students' abilities to work with technology features and accessibility tools;
- Design and development of draft paper-based writing tasks (and supporting documents) for Kindergarten and Grade 1 students, allowing direct assessment of writing skills for students in those grades.

## Table of Contents

Executive Summary .....	i
Acknowledgements.....	1
1. Introduction.....	3
2. Evidence-Centered Design Approach to ELPA21 .....	7
2.1 Domain Analysis.....	9
2.2 Domain Modeling .....	11
2.3 Conceptual Assessment Framework .....	13
2.4 Assessment Implementation .....	14
2.5 Assessment Delivery.....	15
3. ELPA21 Item Design and Development Process.....	16
3.1 Assessment Design Documents .....	16
3.1.1 Test Blueprints .....	16
3.1.2 Reading Level Guidelines.....	17
3.1.3 Proficiency Level Descriptors.....	19
3.2 Task Design .....	20
3.2.1 Item Specifications.....	20
3.2.2 Experimental Items .....	22
3.2.2.A Integrated Tasks .....	22
3.2.2.B More Challenging Reading Passages .....	23
3.2.3 Draft Rubrics for Constructed-Response Tasks .....	23
3.2.4 Draft Scoring Rules for Technology-Enhanced Items .....	25
3.2.5 Item Banking and Metadata .....	26
3.2.5.A Item Banking.....	26
3.2.5.B Metadata .....	27
3.2.6 Editorial Style Guide and Graphics Style Guide.....	28
3.2.6.A Editorial Style Guide.....	29
3.2.6.B Graphics Style Guide.....	29
3.2.7 Sample Items.....	30
3.3 Item Development.....	31
3.3.1 Item Development Plan.....	31
3.3.2 Reading Passage Acquisition and Review .....	33
3.3.2.A Reading Passage Acquisition .....	33
3.3.2.B State Reading Passage Screening Process .....	34

3.3.2.C Educator Passage Review WebEx Meetings.....	34
3.3.3 Item Writing.....	36
3.3.3.A Item Writer Recruitment .....	36
3.3.3.B Development of Item Writer Training Materials.....	36
3.3.3.C Training of ETS Internal and External Contract Item Writers .....	37
3.3.3.D Training of Educator Item Writers .....	38
3.3.4 Item Review .....	41
3.3.4.A Internal ETS Reviews .....	41
3.3.4.B Pre-Committee Meeting Reviews by Task Management Teams .....	43
3.3.4.C Content Review Committee Meetings.....	44
3.3.4.D Bias Review Committee Meetings.....	46
3.3.4.E Process for Handling Issues Outside the Scope of Reviews.....	48
3.3.4.F Results of Content and Bias Review Committee Meetings .....	49
3.3.4.G Resolution WebEx Meetings with Task Management Teams .....	51
3.3.4.H Audio Recording .....	52
3.3.5 Accessibility for Students with Disabilities .....	53
3.3.5.A Employing Principles of Universal Design.....	53
3.3.5.B Collaboration with Accessibility, Accommodations, and Administration Task Management Team.....	54
3.3.5.C Evaluation of Item Accessibility and Use of Accessible Portable Item Protocol.....	55
3.3.5.D Braille-Ready Items for Paper Delivery.....	55
3.3.6 Finalization and Export of Item Pool .....	56
3.3.7 Additional Scope of Work .....	58
3.3.7.A Cognitive Laboratory Study.....	58
3.3.7.B Paper-Based Writing Tasks for Kindergarten and Grade 1 .....	59
References.....	61
APPENDIX A: Additional Tables of Items Produced.....	62
Table 8: Intended Operational Items by Task Type.....	62
Table 9: Field-Test-Ready Items Aligned to Each English Language Proficiency Standard .....	64
Table 10: Field-Test-Ready Items Aligned to Each Practice.....	65
Table 11: Field-Test-Ready Items Aligned to Each Proficiency Level Descriptor (Part 1) .....	66
Table 12: Field-Test-Ready Items Aligned to Each Proficiency Level Descriptor (Part 2) .....	67
Table 13: Sample Items by Task Type.....	68
Table 14: Item Twins and Associated Items .....	70
APPENDIX B: Educator Participants, Demographic Information, and Expertise .....	72

Appendix C: Sources of Data and Quality Control Process Implemented for This Report ..... 83

## Acknowledgements

As the item development contractor for English Language Proficiency Assessment for the 21st Century (ELPA21), Educational Testing Service (ETS) would like to acknowledge the contributions of the many individuals who participated in the design and development of the ELPA21 item pool. We appreciate the collaboration with members of the ELPA21 Consortium and the Council of Chief State School Officers (CCSSO). We wish to acknowledge and thank the following individuals who participated in the item design and development process:

### **ELPA21**

- Bill Auty, Assessment Design and Scaling (ADS) Task Management Team (TMT) Lead
- Wes Bruce, Field Test and Technology Readiness TMT Lead
- Martha Castellon, Executive Director, Understanding Language, Stanford University
- Mark Hansen, Senior Researcher, Center for Research on Evaluation, Standards, and Student Testing, University of California at Los Angeles
- Kara Schlosser, Communications Director
- Mary Seburn, Performance Standard Setting, Data, and Reporting TMT Lead

### **Item Acquisition and Development (IAD) TMT**

- Phoebe Winter, IAD TMT Lead, Assessment Research and Development Consultant
- Michelle McCoy, Education Program Specialist, Oregon Department of Education
- Terri Schuster, Title III Director, ELL Assessment, Nebraska Department of Education
- Kara Todd, Science Assessment Specialist, State of Washington Office of Superintendent of Public Instruction
- Kelly Torres, Instructor, School of Teacher Education at Florida State University

### **Accessibility, Accommodations, and Administration (AAA) TMT**

- Martha Thurlow, AAA TMT Lead, National Center on Educational Outcomes
- Holly Carter, Operations and Policy Analyst, Oregon Department of Education
- Laurene Christensen, Research Associate, National Center on Educational Outcomes
- Andrew Hinkle, Education Program Specialist, Ohio Department of Education
- Nancy Rowch, Director of Language Instruction for Limited English Proficient and Immigrant Students (retired), Nebraska Department of Education
- Vitaliy Shyyan, Research Associate, National Center on Educational Outcomes

### **CCSSO**

- Fen Chou, Program Director, Assessment Standards, Assessment and Accountability
- Margaret Ho, Program Director, ELPA21 Sustainability Planning
- Lauren Lynch, Program Assistant, Assessment Standards, Assessment and Accountability

- Cathryn Still, Program Director, ELPA21

In addition, we would like to recognize the work of the educators from the ELPA21 consortium states who participated in reading passage reviews, wrote items, and served on the content review committee and the bias review committee. The names and affiliations of the individuals who performed this work are listed in Appendix B.

## 1. Introduction

This document summarizes and reports on the activities undertaken by Educational Testing Service (ETS) in 2014 on behalf of and in collaboration with the Council of Chief State School Officers (CCSSO) and the English Language Proficiency Assessment for the 21st Century (ELPA21) Consortium related to the design and development of a pool of test items for the ELPA21 assessment system.

The ELPA21 Consortium is a group of states organized to produce an assessment system to measure the language development of English language learners (ELLs). ELPA21 was awarded a four-year Enhanced Assessment Grant from the U.S. Department of Education in September 2012. ETS was contracted via a Request for Proposal process to complete a scope of work related to the design and development of a pool of field-test-ready test items that would support the ultimate development of an assessment system consisting of a screener assessment and an operational assessment.

The ELPA21 assessment system is intended to be used by educational authorities in the consortium states for:

- Determining initial identification of ELLs (via the screener);
- Monitoring ELLs' annual progress in the attainment of English for academic purposes;
- Measuring districts' success in meeting accountability benchmarks per Title III of the No Child Left Behind act; and
- Consideration for reclassifying students from ELL to Fluent English Proficient status.

ELPA21 is an ambitious undertaking, with the overall goal of improving the English language proficiency assessment of ELL students by bringing to the consortium states an assessment system based on an innovative set of English language proficiency standards (CCSSO, 2013). The following were additional goals for ELPA21 that were central to the item design and development effort:

- The item pool must reflect both the letter and the spirit of the new ELP Standards, including a focus on the English needed for students to communicate and learn grade-appropriate content material in English language arts, mathematics, and science contexts;
- The item pool must take advantage of contemporary approaches to computer-based assessment, with a significant proportion of technology-enhanced (TE) items;
- The item pool must foreground accessibility, ensuring that all test items are maximally accessible to all students, including students with disabilities;
- The item pool must be designed for portability and interoperability, as the item pool produced by ETS would be handed off to another organization for delivery of the field test, and the pool must also be amenable to potential future transitions of the item bank;

- The design and development of the item pool must use Evidence-Centered Design (ECD) to provide an intellectual underpinning that will serve as the basis for the validity argument supporting the ELPA21 assessments.

Several themes related to structuring the work emerged early and persisted throughout the effort. These themes, which proved to be essential to accomplishing the work of designing and developing the ELPA21 item pool on time and with high quality, were:

- The work would need to be done in a very rapid and flexible manner. The federal grant supporting this work had strict timelines requiring that the item pool be delivered by the end of October 2014. This timeline did not support linear processing, with the opportunity for each major task to be completed, reviewed, and confirmed before work began on the next major task. As a result, it was necessary for all stakeholders in this effort to employ non-traditional approaches to work, including substantial parallel processing, and to continually re-think typical approaches to work.
- The work would need to be done in a highly collaborative manner. By design, the organizational structure of the ELPA21 Consortium includes a range of Task Management Teams (TMTs), several of whom had direct stakes in the design and development of the ELPA21 item pool. As detailed throughout this report, the ETS staff interacted and collaborated frequently with the Item Acquisition and Development (IAD) TMT and the Accessibility, Accommodations, and Administration (AAA) TMT. Given the short timelines and the often non-linear work process, these collaborative relationships were crucial.
  - Reinforcing the theme of collaboration, panels of educators from the consortium states played important roles in generating and/or reviewing test materials at several key junctures in the item development process. Educators were positioned to make significant contributions to the ELPA21 item pool because they have expertise in students' language skills and development, have invaluable insight into student interests and grade-appropriate topics, and generally have the ability to strengthen the link between assessment and instruction. As detailed throughout Section 3, panels of educators played key roles in educator passage review WebEx meetings, the item writer training meeting, and the content and bias review committee meetings. A list of those educators who contributed is provided in Appendix B.
- Finally, the conditions listed above required that, to a greater degree than may be typical of assessment design and development efforts, all parties in the current work take an ongoing problem-solving approach. That is, it was not possible to lay out a clear schedule and process in advance that would allow teams to know exactly what work would be accomplished when and how it would be achieved. Instead it was necessary to set major goals and then work quickly, collaboratively, and flexibly to chart a course towards them, addressing myriad complications and challenges that came up along the way.

The major deliverable that ETS was responsible for was the design and development of the field test pool for the ELPA21 assessment system. As detailed in Table 1 and Table 2, the pool contained a total of 2,619 field-test-ready items. This full set of field-test-ready items was made up of two sub-categories: 2,469 intended operational items (i.e., produced with the intention that, once the quality of the items is confirmed via field testing, they will make up the initial operational forms) and 150 experimental items (i.e., produced to be field tested but not intended for use in initial operational forms).<sup>1</sup> Of the intended operational items, 1,138 or 46 percent are TE items and 1,076 or 44 percent contain accessible content authored via Accessible Portable Item Protocol (APIP).

**Table 1: Field-Test-Ready Items by Grade or Grade Band and Domain**

<b>Grade or Grade Band</b>	<b>Listening</b>	<b>Reading</b>	<b>Speaking</b>	<b>Writing</b>	<b>Total</b>
<b>K</b>	189	145	123	91	<b>548</b>
<b>1</b>	163	156	75	75	<b>469</b>
<b>2-3</b>	129	150	65	93	<b>437</b>
<b>4-5</b>	133	136	85	72	<b>426</b>
<b>6-8</b>	127	122	53	39	<b>341</b>
<b>9-12</b>	127	170	54	47	<b>398</b>
<b>Total</b>	<b>868</b>	<b>879</b>	<b>455</b>	<b>417</b>	<b>2619</b>

**Table 2: Intended Operational Items by Item Type/Response Format: Constructed-Response (CR); Selected-Response (SR); and Technology-Enhanced (TE) Items**

<b>Grade or Grade Band</b>	<b>CR</b>	<b>SR</b>	<b>TE</b>	<b>Total</b>
<b>K</b>	123	63	342	528
<b>1</b>	75	108	246	429
<b>2-3</b>	105	124	181	410
<b>4-5</b>	113	137	157	407
<b>6-8</b>	83	156	83	322
<b>9-12</b>	85	159	129	373
<b>Total</b>	<b>584</b>	<b>747</b>	<b>1138</b>	<b>2469</b>

In addition to the item pool itself, ETS also produced the following deliverables:

- Completion of Stage 1 Test Blueprints, providing information about anticipated test instances, item response types, and score points for the ELPA21 summative assessment (building on work begun by the TMTs);
- Development of Reading Level Guidelines describing the text complexity of ELPA21 reading passages;

<sup>1</sup> More detailed information on experimental items is provided in Section 3.2.2.

- Completion of draft Proficiency Level Descriptors (PLDs) specific to each standard and domain (building on work begun by the TMTs);
- Stage 2 Item Specifications (one for each of the ELPA21 grades or grade bands: Kindergarten, Grade 1, Grades 2-3, Grades 4-5, Grades 6-8, and Grades 9-12), providing detailed information about the design of all task types developed for ELPA21;
- Draft rubrics for constructed-response (CR) task types;
- Draft scoring rules for TE task types, including draft partial credit scoring rules for those TE task types judged to be amenable to the awarding of partial credit;
- Definition of the metadata fields and available values to be assigned to and associated with each test item;
- An Editorial Style Guide and a Graphics Style Guide, developed to help ensure that presentation of language and images for each grade or grade band is consistent, effective, and grade appropriate;
- A pool of 193 sample items provided as an element of the Item Specifications in order to guide item writers and also delivered in the same computer-based format as the field test pool to allow for potential use for test familiarization purposes;
- A supplementary pool of 415 “twin” items designed to be accessible to students with visual impairments. These items are designed to support development of paper-based test forms that will generate scores comparable to the computer-delivered forms;
- Design and execution of a cognitive laboratory study to assess how well students can interact with various computer-based item types, the clarity of the directions, and students’ abilities to work with technology features and accessibility tools;
- Design and development of draft paper-based writing tasks (and supporting documents) for Kindergarten and Grade 1 students, allowing direct assessment of writing skills for students in those grades.

## 2. Evidence-Centered Design Approach to ELPA21

As mentioned in Section 1, one of the key guiding principles for the development of the ELPA21 assessment system is to employ an ECD approach. This section provides a brief overview of ECD and describes how each layer of ECD is represented in the development of the ELPA21 assessment system.

ECD looks on an educational assessment as “an evidentiary argument for reasoning what students say, do, or make in particular task situations as well as to generally claim what they can know, do, or have accomplished” (Mislevy, 2011, p. 6). From this perspective, ECD provides a “principled framework” (Mislevy, Steinberg, & Almond, 2003, p. 1) for constructing and documenting such an evidentiary argument through the process of conceptualizing, designing, developing, implementing, and operating an educational assessment such as ELPA21.

ECD extends evidence of what students do in a testing situation to empirically derived claims about what they know and can do in the real world. For the purposes of assessment design, ECD is not a rigid set of procedures for developing items and tasks; rather, it is a set of principles and procedures for assessment design and development that require clear and specific documentation of:

- The purpose of the assessment;
- The claims to be made about student performance;
- The knowledge, skills, and other attributes that the assessment is to measure;
- The relationship between the knowledge and skills to be measured and those which the items and tasks, supported by scoring materials, actually measure;
- The relationship between student performance on individual items and aggregate performance on the assessment to overall conclusions to be made about student performance.

Although the ECD approach is inherently iterative, it is commonly conceptualized as a series of five layers that constitute a progression from more abstract conceptualization to more specific and concrete instantiation. Table 3 presents the five layers of ECD (domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery) along with the corresponding elements associated with each layer for ELPA21. Work products that were created (either by ETS or by ELPA21) as part of the item pool design and development, which is the basis of this report, are shown in **bold**. Each of these layers, and their instantiation for ELPA21, is briefly described after the table.

**Table 3: Realization of Evidence-Centered Design Layers via ELPA21 Elements**

<b>Layers of ECD</b> (adapted from Mislevy, 2011)		<b>Corresponding ELPA21 Elements</b>
Domain analysis	<ul style="list-style-type: none"> <li>• What is important about this domain?</li> <li>• What work and situations are central in this domain?</li> <li>• What knowledge representations are central?</li> </ul>	<ul style="list-style-type: none"> <li>• English Language Proficiency (ELP) Standards (CCSSO, 2014)</li> <li>• English Language Proficiency Development (ELPD) Framework (CCSSO, 2012)</li> </ul>
Domain modeling	<ul style="list-style-type: none"> <li>• How do we articulate the assessment argument based on the domain analysis?</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Claims</b></li> <li>• <b>Sub-claims</b></li> <li>• <b>PLDs</b></li> </ul>
Conceptual assessment framework	<ul style="list-style-type: none"> <li>• How do we coordinate the substantive, statistical, and operational aspects of the assessment?</li> </ul> Design structures: <ul style="list-style-type: none"> <li>• Student model</li> <li>• Evidence model</li> <li>• Task model</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Stage 1 Test Blueprints</b></li> <li>• <b>Item Specifications</b></li> <li>• <b>Sample items</b></li> <li>• <b>Reading Level Guidelines</b></li> <li>• <b>Editorial Style Guide</b></li> <li>• <b>Graphics Style Guide</b></li> <li>• <b>Draft rubrics for CR items</b></li> <li>• <b>Draft scoring rules for TE items</b></li> <li>• <b>Metadata fields</b></li> </ul>
Assessment implementation	Production aspects of assessment development: authoring tasks, scoring details, statistical models	<ul style="list-style-type: none"> <li>• <b>Item Pool and Its Supplements:</b> <ul style="list-style-type: none"> <li>○ <b>Intended operational items</b></li> <li>○ <b>Experimental items</b></li> <li>○ <b>Twin items</b></li> <li>○ <b>Grades K and 1 paper-based writing tasks</b></li> </ul> </li> </ul>
Assessment delivery	Students interact with tasks, performances evaluated, feedback created.	<ul style="list-style-type: none"> <li>• Field test (and revisions/refinements based on field test results)</li> <li>• Operational administrations and operational score reporting</li> </ul>

## 2.1 Domain Analysis

ELPA21 is an assessment system concerned with the English language proficiency of Kindergarten to Grade 12 students in the ELPA21 consortium states. The real-world domains of interest, which are key considerations in ECD, are these students' developing English language skills as they relate to the academic content knowledge required for students to be college- and career-ready by the end of high school.

The functional domain analysis for ELPA21—the documentation of these real-world domains in terms that support assessment design and development (as well as other purposes such as development of curricula and instruction)—is contained in the English Language Proficiency (ELP) Standards (CCSSO, 2014).

The ELP Standards are organized into six grades or grade bands (Kindergarten, Grade 1, Grades 2-3, Grades 4-5, Grades 6-8, Grades 9-12), and for each grade or grade band the standards “highlight and amplify the critical language, knowledge about language, and skills using language that are in college-and-career-ready standards and that are necessary for English language learners to be successful in schools” (p. 1).

As the key overall guiding document for the ELPA21 item pool design and development work, the ELP Standards are the basis for the conceptualization of language proficiency, the conceptualization of student progress (or sequence of language development), and, at the most general level, the organizational structure (e.g., the organization into the six grades or grade bands) of the ELPA21 assessments.

The ELP Standards emphasize the importance of recognizing that, particularly in the context of the Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS), students engage actively in learning in the content areas even as their English language proficiency progresses. This emphasis is borne out by the focus that the ELP Standards place on correspondences with the CCSS and the NGSS, as instantiated by “practices,” which the ELP Standards define as “behaviors which developing student practitioners should increasingly use when engaging with the content and growing in content-area maturity and expertise” (p. 31). The practices are the key means through which the standards encode the emphasis on communicative skills in the academic content areas as a central element of the definition of the domain of interest for ELPA21. The ELP Standards on which ELPA21 is based are presented in Table 4.

**Table 4: Organization of the English Language Proficiency Standards in Relation to Participation in Content-Area Practices**

1	construct meaning from oral presentations and literary and informational text through grade-appropriate listening, reading, and viewing
2	participate in grade-appropriate oral and written exchanges of information, ideas, and analyses, responding to peer, audience, or reader comments and questions
3	speak and write about grade-appropriate complex literary and informational texts and topics
4	construct grade-appropriate oral and written claims and support them with reasoning and evidence
5	conduct research and evaluate and communicate findings to answer questions or solve problems
6	analyze and critique the arguments of others orally and in writing
7	adapt language choices to purpose, task, and audience when speaking and writing
8	determine the meaning of words and phrases in oral presentations and literary and informational text
9	create clear and coherent grade-appropriate speech and text
10	make accurate use of standard English to communicate in grade-appropriate speech and writing

Standards 1 through 7 involve the language necessary for ELLs to engage in the central content-specific practices associated with ELA & Literacy, mathematics, and science. They begin with a focus on extraction of meaning and then progress to engagement in these practices.

Standards 8 through 10 hone in on some of the more micro-level linguistic features that are undoubtedly important to focus on, but only in the service of the other seven standards.

(CCSSO, 2014, p. 4)

Note that the standards are highly integrated in nature and several of them cross the domains of listening, reading, speaking, and writing. The same ten standards are established for each of the six grades or grade bands.

For each of the ten standards at each of the six grades or grade bands, the ELP Standards also provide descriptors for five distinct proficiency levels, which provide information about what typical student language proficiency “looks like” as ELLs progress toward meeting each standard.

The standards themselves cite (and rely fairly heavily upon) the *Framework for English Language Proficiency Development Standards Corresponding to the Common Core State Standards and the Next Generation Science Standards*, commonly referred to as the “ELPD Framework” (CCSSO, 2012). This document contains the theoretical underpinnings on which the ELP Standards are based, including an emphasis on “a reconceptualization of the way English Language Learners (ELLs) ‘apprentice’ into [the] demanding disciplinary practices” represented by the CCSS and the NGSS by “simultaneously acquiring and developing language as well as acquiring disciplinary knowledge and skills” (p. 1).

A careful analysis and understanding of both the ELP Standards and the ELPD Framework were crucial to the subsequent work on the ELPA21 item pool, particularly in ensuring that the strong

emphases on the conceptual understanding of language proficiency, the importance of practices, and the conceptualization of student progress were embodied in the later stages of item design and development.

## 2.2 Domain Modeling

The ELP Standards have as their primary focus the definition of English language proficiency as needed to inform curriculum and instruction. To develop the ELPA21, it is necessary to articulate a principled manner of interpreting and sampling this proficiency so it can be measured within the confines of a standardized assessment with practical time limits. Although several of the standards call for the integration of skills, Title III mandates that students be assessed in the four separate domains of listening, reading, speaking, and writing. Therefore, as the assessment domain was modeled, it was necessary to begin by articulating the assessment in terms of four separate skill domains as they relate to the ELP Standards. This was a crucial decision in the domain modeling stage. While recognizing the multidimensional nature of English language development and the emphasis on collaborative skills in the ELP Standards, ELPA21 would rely on test items designed to measure skills by domain (i.e., listening, reading, speaking, writing). This decision was discussed in an issue brief on multidimensionality issued by the IAD TMT. Three documents were later developed that define and document this decision: the ELPA21 claims, the ELPA21 sub-claims, and the disaggregated ELPA21 PLDs.

Claims: The high-level ELPA21 claims, which are domain-level statements about student abilities, are shown below.

- The English language learner can listen and comprehend spoken English in the context of grade-appropriate activities.
- The English language learner can read and comprehend written English in the context of grade-appropriate activities.
- The English language learner can produce comprehensible speech that is typical of grade-appropriate activities.
- The English language learner can write comprehensible texts that are the result of grade-appropriate activities.

Sub-Claims: The ELPA21 sub-claims represent a disaggregation of the ten ELP Standards across the four domains of listening, reading, speaking, and writing. For example, Standard 3, “An ELL can speak and write about grade-appropriate complex literary texts and topics” relates to both speaking and writing skills. For the purposes of the assessment, however, it was necessary to disaggregate Standard 3 into a speaking claim and a writing claim. The examples below show how Standard 3 maps on to sub-claims for speaking (3S) and writing (3W):

3S The English language learner speaks about complex literary and informational texts and topics.

3W The English language learner writes about complex literary and informational texts and topics.

Because not all claims are relevant to all of the four domains (for example, Standard 3, which focuses on productive skills, maps on to sub-claims for speaking and writing but not the receptive skills of listening or reading) there are a total of 26 sub-claims.

Proficiency Level Descriptors: The ELPA21 Standards contain PLDs for each of the ten ELP Standards. The PLDs describe targets for ELL performance by the end of each ELP level. While the ten standards are consistent across all grades, the PLDS for each standard are different for each grade or grade band. That is, the standard is further defined with grade-appropriate expectations at each of the five proficiency levels.

In order to relate the PLDs to the domain-specific assessment, the PLDs, like the standards, were also disaggregated. Below are the PLDs for ELP Standard 3 for Grades 4-5.

ELP Standard		By the end of each English language proficiency level, an ELL can . . .				
		Level 1	Level 2	Level 3	Level 4	Level 5
4-5.3	An ELL can . . .			including a few details,	including some details,	including details and examples to develop a topic,
	speak and write about grade-appropriate complex literary and informational texts and topics.	<ul style="list-style-type: none"> <li>communicate simple information</li> </ul> about familiar texts, topics, events, or objects in the environment.	<ul style="list-style-type: none"> <li>deliver short oral presentations</li> <li>compose written texts</li> </ul> about familiar texts, topics, and experiences.	<ul style="list-style-type: none"> <li>deliver short oral presentations</li> <li>compose written narratives or informational texts</li> </ul> about familiar texts, topics, and experiences.	<ul style="list-style-type: none"> <li>deliver short oral presentations</li> <li>compose written narratives or informational texts</li> </ul> about a variety of texts, topics, and experiences.	<ul style="list-style-type: none"> <li>deliver oral presentations</li> <li>compose written narrative or informational texts</li> </ul> about a variety of texts, topics, and experiences.

The domain-specific PLDs for Standard 3, disaggregated to show specific application to speaking and to writing, respectively, are shown below.

ELP Standard 3					
Speaking Sub-claim 3S: The English language learner speaks about complex literary and informational texts and topics. The learner can:					
Level 1	Level 2	Level 3	Level 4	Level 5	
<ul style="list-style-type: none"> <li>communicate simple information about familiar texts, topics, events, or objects</li> </ul>	<ul style="list-style-type: none"> <li>deliver short oral presentations about familiar texts, topics, and experiences</li> </ul>	including a few details, <ul style="list-style-type: none"> <li>deliver short oral presentations about familiar texts, topics, and experiences</li> </ul>	including some details, <ul style="list-style-type: none"> <li>deliver short oral presentations about a variety of texts, topics, and experiences</li> </ul>	including details and examples to develop a topic, <ul style="list-style-type: none"> <li>deliver oral presentations about a variety of</li> </ul>	

in the environment				texts, topics, and experiences
--------------------	--	--	--	--------------------------------

<p><b>ELP Standard 3</b>  <b>Writing Sub-claim 3W:</b> The English language learner writes about complex literary and informational texts and topics. The learner can:</p>				
Level 1	Level 2	Level 3	Level 4	Level 5
<ul style="list-style-type: none"> <li>communicate simple information about familiar texts, topics, events, or objects in the environment</li> </ul>	<ul style="list-style-type: none"> <li>compose written texts about familiar texts, topics, and experiences</li> </ul>	<p>including a few details,</p> <ul style="list-style-type: none"> <li>compose written narratives or informational texts about familiar texts, topics, and experiences</li> </ul>	<p>including some details,</p> <ul style="list-style-type: none"> <li>compose written narratives or informational texts about a variety of texts, topics, and experiences</li> </ul>	<p>including details and examples to develop a topic,</p> <ul style="list-style-type: none"> <li>compose written narrative or informational texts about a variety of texts, topics, and experiences</li> </ul>

Although the field test pool consisted of items meant to measure language proficiency by individual domain, the ELPA21 Consortium retains an active interest in the potential for developing test items that are more multidimensional in nature. To support possible future work in this direction, some experimental items measuring skills across domains were included in the field test pool (see details in Section 3.2.1).

### 2.3 Conceptual Assessment Framework

This layer of ECD—consisting of a student model, a task model, and an evidence model bridging the two—is the first of two layers in which the great majority of the work done on the ELPA21 item-pool design and development took place. Because this work is described in considerable detail in Section 3, it will be discussed relatively briefly here.

The student model consists of a representation of the knowledge, skills, and abilities of students who will be taking ELPA21 with respect to the construct of interest: English language proficiency in the domains of listening, reading, speaking, and writing. Introductory text for each domain within the Item Specifications provides a high-level description of how the construct is understood and interpreted for purposes of assessment. The sub-claims, which each task has been designed to gather evidence to support, are also listed for each task, further defining the construct for each domain. Additionally, at a more granular level, the ELPA21 PLDs (as disaggregated for each domain) serve as a representation of expected abilities of typical students across five levels of proficiency for each of the standards.

The key ELPA21 elements for the task model are the Item Specifications and the Test Blueprints. The Item Specifications describe in detail the various task types that will be included on ELPA21, documenting both fixed elements (those that are present in all tasks of a given type) and variable elements (those that differ and provide each individual task with its own particular

qualities of measurement). The Item Specifications are a key driver of generativity for the ELPA21 because they help to ensure that all of the tasks in the item pool have an appropriate range of similarities and differences to capture the evidence needed to serve the intended purposes of the assessment. In the case of ELPA21, the Item Specifications were also a key point of early review and discussion to ensure that all key stakeholders, particularly the TMTs and the ETS Team, had similar expectations for what was to be produced as the items for the pool were drafted and reviewed. The Test Blueprints provide information about how the task types described in the Item Specifications will be assembled into test forms<sup>2</sup>; the number of tasks of each type to be included at each grade or grade band; and the number of score points to be generated from selected-response (SR) items, from TE items, from CR items, and in total.

An evidence model consists of two components, an evaluation component and a measurement model; the first of these was substantially completed as part of the item pool design and development effort, while the latter was beyond the scope of this work. The evaluation component refers to how individual tasks are to be scored; for ELPA21 this consists of identified keys (i.e., correct answers) for SR tasks, draft scoring rules for TE items, and draft rubrics for CR items. For SR items, the key is identified and verified as each item is drafted and reviewed. The processes for developing the draft rubrics for CR items and the draft scoring rules for TE items are described in sections 3.2.3 and 3.2.4, respectively.

The rubrics and scoring rules are described as “draft” at this point because they are subject to validation and refinement at later stages of the assessment design process, specifically, the point at which they can be evaluated in light of their effectiveness in operation. This will occur once actual student responses to tasks have been generated via the field test.

## 2.4 Assessment Implementation

This layer is the point in the ECD process at which the assessment design and development process moves from the conceptual to the more practical. Given the iterative nature of assessment design, many elements of the conceptual assessment framework (e.g., the rubrics, the automated scoring rules) were reviewed and refined in this layer. Because the focus of the work described in this report is on the design and development of an item pool, the largest and most significant work product for ELPA21 at this stage was the writing and reviewing of the numerous test items and tasks that make up the item pool itself.

The development of this pool, including the key roles played by a range of ELPA21 stakeholders, is described in detail in Section 3. The pool consists of a large number of test items intended for use with the general ELPA21 student population; a set of experimental items that do not strictly conform to the ELPA21 task model but were developed for research purposes; a set of “twin items” developed to support the development of accessible forms for students with

---

<sup>2</sup> While item development was based on the construction of a specified number of linear test forms, the consortium plans to transition to adaptive testing.

visual impairments; and a set of draft writing items for Kindergarten and Grade 1 students designed to be administered in a paper-based form to complement those writing tasks included in the larger, computer-based ELPA21 item pool at those grades. In addition, a Cognitive Laboratory study was designed and implemented, providing feedback about student responses to selected task types, directions, and accommodation features; results can be used to refine the delivery of the ELPA21 items.

## **2.5 Assessment Delivery**

The final layer of ECD is beyond the scope of the item-pool design and development effort described in this report. ELPA21 will have two major assessment delivery events: the field test administration and the operational administration. In the field test, the key steps from an ECD perspective will be assembly of field test forms, administration of those forms, item-level scoring, analysis and interpretation of field test results, and revisions to elements of the conceptual assessment framework and assessment implementation documents as appropriate based on those field test results. When turning to operational administration, the key ECD steps will be assembly of operational forms, administration of those operational forms, item-level scoring, and generation of reported scores.

### 3. ELPA21 Item Design and Development Process

#### 3.1 Assessment Design Documents

##### 3.1.1 Test Blueprints

The ELPA21 Test Blueprints are documents, organized by domain, which define what each test form (or, in language sometimes used for computer-based assessment, “test instance”) will contain. Much as the blueprint for a building provides a description of the shape and key elements in advance of construction, the ELPA21 Test Blueprints serve as guiding documents for the item development by ensuring that an appropriate number and distribution of items are developed in order to serve for the later assembly of the needed test forms for the ELPA21 for each grade band. Specifically, they served as a guide for how to populate the field test pool with enough overage to allow for two full operational forms and a screener.

Developing the Test Blueprints was a collaborative and iterative process that could be described as an effort to optimize fulfillment of the requirements within a series of constraints. A key requirement was that ELPA21 be able to report scores for each domain (listening, reading, speaking, and writing). A key constraint was that of testing time. The process began with targets of assessments that would be limited to one hour for Kindergarten and for Grade 1, one and a half hours for Grades 2-3, and two hours for Grades 4-5, Grades 6-8, and Grades 9-12.

Given that ELPA21 contains ten largely integrated standards and seeks to report scores for four separate domains, it was necessary to evaluate the possible set of assessment task types and select a set of task types that would 1) produce a reliable score for each domain, 2) cover a range of standards for each domain, 3) cover a standard adequately (i.e., not overemphasizing a standard by assessing it across all four domains, but not assessing it only in one domain if sub-claims are spread across all four domains), and 4) gather evidence of proficiency for students at different levels including students who are at very low levels of proficiency and those who are at advanced levels of proficiency. Selecting a set of final tasks for inclusion also involved the need to make judgments about the feasibility of administering a set of tasks in a large-scale assessment setting without proving to be a burden to students. In addition, there was a need to balance efficient-to-score SR items with direct measures of speaking and writing performance via CR items and to consider ways to limit the burden (i.e., cost and time) of scoring. There was a general emphasis on using TE items to the degree appropriate, given the importance placed on TE items in the Enhanced Assessment Grant that provided the funding for ELPA21.

Several initial formats for Test Blueprints were explored and served as the basis of productive conversations among the IAD TMT, the Assessment Design and Scaling (ADS) TMT, and the ETS Team; it became clear over time that the selection of task types would drive the finalization

of the Test Blueprints, and the final drafts of the Test Blueprints were structured around task types.

Stage 1 Test Blueprints, which served as the basis for the Item Development Plan, were developed by ETS and organized by domain, reflecting the requirement to report scores by domain. Within each domain, the Test Blueprints detail the number of items to be included on test forms at each grade or grade band as distributed across task types and response formats (SR, TE, short CR, and extended CR). The Test Blueprints also show how many total items each test form will contain and an estimate of how many score points will be generated for each grade or grade band in each domain. The initial drafts of these Test Blueprints were reviewed extensively by the IAD TMT and ADS TMT, and a number of adjustments were made both to ensure that an appropriate number of score points were provided for each domain at each grade or grade band and for clarity of presentation. Once the TMTs were satisfied with the content and presentation of the Test Blueprints, they were reviewed and approved by the ELPA21 Consortium Council.

While the Stage 1 Test Blueprints served as the basis of the development of the ELPA21 item pool, it is expected that the Test Blueprints will be revised to provide, for example, more detailed plans about the structure of the screener assessment (for Stage 1, the conservative estimate was made that the screener will be the same length as the summative assessment), and to reflect numbers of score points associated with CR and TE tasks after their rubrics and scoring rules have been validated based on field test results.

### *3.1.2 Reading Level Guidelines*

The purpose of the ELPA21 Reading Level Guidelines was to standardize a definition of reading “level” so that suitable texts could be obtained for use as reading passages. A key decision informing the Reading Level Guidelines was that the most challenging reading passages in each ELPA21 grade band would represent the level of complexity represented by the CCSS expectations for all students at the lowest grade within the grade band (e.g., for Grades 6-8, the most complex texts would be at a 6<sup>th</sup> grade level).

For Grades 2 and above, the complexity of ELPA21 passages were evaluated by TextEvaluator™, an automated readability tool developed at ETS to help teachers, textbook publishers, test developers, and literacy researchers select reading materials that are consistent with the text complexity goals outlined in the CCSS. For Kindergarten and Grade 1, levels at which TextEvaluator scores are less reliable, appropriate passage complexity was determined by the professional judgment of ETS test developers and educators who participated in the passage review WebEx meetings and confirmed by educators on the content review committee. Table 5 shows the TextEvaluator scores for Grades 2-12 reading passages as finalized after the content review committee meetings, along with the number of items associated with the passages.

**Table 5: Items Associated with Passages by TextEvaluator Text Complexity Score**

<b>Text Complexity</b>	<b>2-3</b>	<b>4-5</b>	<b>6-8</b>	<b>9-12</b>	<b>Total</b>
3	28	4			32
4	3				3
7	3				3
8	7	3			10
9	4				4
11	17				17
12		4			4
13	3	4			7
14	1				1
16	4				4
18	10				10
19	1				1
20		7	4		11
21		4			4
22		4			4
23		4			4
24	4	9			13
25		10			10
26		4			4
27	4				4
28		5	4		9
29		14			14
31	7	4	11		22
32			3		3
33			4		4
34		4			4
35			10		10
36			9		9
37			9		9
38		4	8	4	16
39			7		7
40				11	11
41			4	13	17
43			4	6	10
45		5		10	15
46			5	6	11
47				5	5
48			4		4
49				12	12
53				3	3
54				4	4

<b>Text Complexity</b>	<b>2-3</b>	<b>4-5</b>	<b>6-8</b>	<b>9-12</b>	<b>Total</b>
56				11	11
58				12	12
59				5	5
61				5	5
66				4	4
69				5	5
<b>Total</b>	<b>96</b>	<b>93</b>	<b>86</b>	<b>116</b>	<b>391</b>

A feature of the TextEvaluator tool is that text complexity for each grade band is represented by a range of scores, and these scores overlap across grades. For example, a text with a TextEvaluator score of 33 (on a 100-point scale) is at the top of the targeted complexity level for Grades 4-5 and is near the middle of targeted complexity level for Grades 6-8.

While the principles outlined in the Reading Level Guidelines and the TextEvaluator scores assigned to passages were valuable input as the ELPA21 passages were acquired and reviewed, TextEvaluator scores were not given priority over judgments by educators and other professionals. For example, content review committees made several small adjustments to reading passages<sup>3</sup>, generally intended to enhance the clarity of the passages, which resulted in a small number of TextEvaluator scores being slightly outside of the planned range. These adjustments by educators were reflected in the final version of the items despite their impact on the TextEvaluator scores, as ELPA21 had made the decision to prioritize professional judgment over TextEvaluator scores.

The Reading Level Guidelines delivered under the current scope of work are labeled as Stage 1. ETS recommends that these Stage 1 guidelines be reviewed and either confirmed or adjusted after empirical results of the field test are available.

### *3.1.3 Proficiency Level Descriptors*

As described briefly in Section 2, two types of PLDs have been developed for ELPA21. The ELP Standards contain descriptors of five levels of proficiency for each of the ten standards at each grade or grade band. Given the requirement to report scores for the four distinct domains, it became necessary to disaggregate these descriptors to provide specific information about typical levels of student performance for each standard within listening, reading, speaking, and writing.

This disaggregation was essentially an administrative task of separating out, for example, the statements made for Standard 1 about listening and reading into distinct statements, one for listening and another for reading. The IAD TMT began this work and ETS completed the disaggregation following the model established by the IAD TMT. The disaggregated PLDs were

<sup>3</sup> The ELPA21 item pool contains no copyrighted texts; as a result, such minor edits could be made.

included in the Item Specifications for each grade or grade band. The disaggregated, domain-specific PLDs became an important element of the metadata for each item as described in the Item Specifications and as reviewed by educators at the content review committee meetings.

## 3.2 Task Design

### 3.2.1 Item Specifications

The Item Specifications occupied a place of great importance in the ELPA21 item pool design and development effort because they document and describe all of the qualities of each task type to be contained in ELPA21. The Item Specifications were the basis of all item development activities, including the development of the pool of field-test-ready items, and should act as the basis for potential future item development. The Item Specifications also served as a crucial point of review by ELPA21 TMTs in establishing shared expectations and understandings among those stakeholders and ETS about the work on which we would collaborate during the production stages of item writing and review. Finally, the Item Specifications were, to a considerable degree, developed hand-in-hand with the Test Blueprints as each of these documents were highly interdependent.

Six Item Specifications documents were developed for ELPA21, one for each of the grades or grade bands. Within each set of Item Specifications, separate specifications are provided for the four domains of listening, reading, speaking, and writing.

For each domain, the Item Specifications begin with the ELPA21 high-level claim and with an operational definition of how the construct was defined to gather evidence about that claim. Then an overview is provided of the task types included in this domain, the various stimuli, the response formats, and the scoring approaches. The introductory matter of each set of Item Specifications ends with a table listing the task types and the standards each is designed to assess.

The bulk of each set of Item Specifications consists of detailed descriptions of each task type for that grade or grade band and domain. These descriptions, which served as guidelines for item writers, include an overall description of the task type; key information about presentation (the directions, the characteristics of the stimulus, etc.); notes on accessibility; and the claims and sub-claims the task type was designed to assess, including specific proficiency levels. The detailed descriptions are followed by a selection of sample items that illustrate the task type.

The process of developing the Item Specifications was a highly iterative and collaborative one. The ETS Team began work, after a thorough review of the ELP Standards, by reviewing the 26 sub-claims and brainstorming tasks types that might be appropriate in gathering evidence of student proficiency related to one or more of them. At this point, the ETS Team considered

known task types (including a number of innovative task types developed by ETS via a recently completed internal research project to advance innovative assessment of K-12 ELLs), adaptations of known task types, and novel task types created in response to the ELP Standards. Given that the ELP Standards are the same across all grades or grade band, the ETS Team looked for task types that would work effectively at a range of grades to support consistency and vertical articulation across ELPA21 with the understanding that while standards are the same across grades, the topics, text complexity, and expectations of students would differ, as articulated in the PLDs for each grade or grade band.

Once a pool of task types had been proposed, they were reviewed and evaluated against a range of criteria including:

- Effectiveness in eliciting evidence for the identified sub-claim(s), including number of possible score points (with TE and CR items likely to be worth multiple score points);
- Feasibility of administration in a standardized testing context;
- Ease of replicability (with appropriate variations) given the number of tasks to be produced;
- Compatibility with APIP standards
- Overlap with other task types under consideration;
- Number of task types assessing a given standard;
- Number of task types needed for a given domain;
- Number of overall task types.

(See Table 8: Intended Operational Items by Task Type in Appendix A for information about the number task types included in each grade or grade span.)

At this point, the work on selecting task types interacted directly with finalization of the Test Blueprints, as described in Section 3.1.1 above. Once a final set of decisions had been made as to which task types would be developed (as documented in the Test Blueprints), work began on drafting and reviewing the Item Specifications.

The ETS Team began by drafting initial versions of the Item Specifications (including sample items in manuscript form) for Grade 1 and Grades 6-8. This approach was chosen because it allowed the IAD TMT to provide an early review of the general approach to Item Specifications before the documents for all six grades and grade bands were developed. The IAD TMT provided extensive feedback and discussion on this early draft, including input on related topics such as the directions for several task types, the appropriateness of including elements of fantasy in task types for the younger grades, and the importance of developing draft rubrics along with the Item Specifications. The IAD TMT also requested that ETS produce functional (i.e., computer-based) versions of the sample items for review along with the Item Specifications. (See Section 3.2.7 for more information on the sample items.)

After this initial review, the ETS Team produced complete draft versions of the Item Specifications for all the grades and grade bands for review by the IAD TMT, along with draft

rubrics for CR items and functional sample items, which the IAD TMT reviewed in the Item Banking and Information System (IBIS) database. This round of review and the changes made by the ETS Team based on it led to the Stage 1 Item Specifications, which served as the basis for item writer training and the development of the item pool. However, the discussions made clear that several policy decisions would need to be made before the Item Specifications could be finalized (e.g., whether students would be allowed to replay listening stimuli).

Over the course of the item writing and review process, the ETS Team continued to take notes for the further refinement of the Item Specifications. These notes reflected substantive decisions made based on input from educators at the content and bias review committee meetings, such as the decision of the Kindergarten review panel that all reading stimuli should be read aloud; the decision of the Grade 1 panel to drop a TE writing task type found not to align to standards (“Complete the Story”); the decision of the Grades 2-3 panel to revise a TE writing task type (“Word Builder”) to better align to standards; more detailed information about accessibility practices for students with visual impairments; and logistical changes such as further standardization of directions that was implemented during the audio recording process. Once the item development effort had been completed, the ETS Team produced Stage 2 Item Specifications as part of the final deliverables for the current contract.

### *3.2.2 Experimental Items*

The focus of the Item Specifications documents was to define expectations for those items that were intended for possible use, after field testing, in the initial operational forms of ELPA21. However, the field test also offered an opportunity to develop and evaluate some task types that were not intended to be used in the initial operational forms but which might provide information about potential new task types to be considered for use on future versions of ELPA21. At the request of the IAD TMT, ETS developed two such categories of items: integrated tasks and more challenging reading passages.

#### *3.2.2.A Integrated Tasks*

While the ELP Standards are highly integrated, as mentioned previously, the ELPA21 assessments seek to limit each task to target a single domain. To explore the potential appropriateness of multidimensional or integrated skills task types, ETS developed an additional set of CR items based on existing listening or reading passages. These CR items were added to the SR and/or TE items included in the “non-experimental” versions of these sets.

These new CR items are distributed as follows: In Kindergarten and Grade 1, four new short CR speaking items per grade; in the higher grade bands, two new extended CR speaking items and two new extended CR writing items per grade band. The sets containing these items are coded as “Experimental Items-Integrated.” While these items are intended to be field tested, they are not intended to be used operationally unless the test specifications are updated to call for such integrated tasks.

### *3.2.2.B More Challenging Reading Passages*

As described in Section 3.1.2, a preliminary guideline established for ELPA21 reading passages is that the most challenging passages in each grade band will represent the level of complexity represented by Common Core expectations for all students at the lowest grade level within the grade band. The test specifications established the upper limit of text complexity score ranges at the mid-point for the lowest grade in the band to avoid requiring students in the lower grades in the band to read passages that were above expectations for their current grade.

To explore the appropriateness of more challenging reading passages for each grade band, including passages at a complexity reaching into the range of the highest grade level in the band, ETS developed several reading passages at higher than the established levels of text complexity for grades 2-12. These passages were not originally intended for use on initial operational forms and were intended to be coded as “experimental” items, analogous to the coding of the integrated items described above.

All of the more challenging passages were reviewed and deemed appropriate for each grade band by the passage review committee, and later the passages and items were also judged to be appropriate for each grade band. In other words, although the TextEvaluator scores for some passages were above the established range for a grade band, teachers judged the complexity and general level of challenge to be acceptable for the targeted grade band.

It should also be noted that the complexity of passages evolved as the passages were reviewed and revised by the passage review panels, educator item writers, content review committees, and bias review committees. Due to the revisions, the complexity of some passages changed (both increasing and decreasing); as a result, the passages originally developed to be “more complex” did not remain as the most complex passages in the grade band. In the end, the following numbers of reading passages above the targeted TextEvaluator levels were included in the pool: three passages in Grades 2-3; three passages in Grades 4-5; one passage in Grades 6-8; and two passages in Grades 9-12.

Because of the lack of a clear line of demarcation between passages intended as “more complex” and the remainder of the passages, passages intended to reflect higher difficulty levels have not been coded as experimental items. ETS’s recommendation is that field test results be used to determine which passages are appropriate for use on initial operational forms and which are not.

### *3.2.3 Draft Rubrics for Constructed-Response Tasks*

For all CR tasks, those which directly measure speaking and writing, a key part of the task design and development process is the creation of scoring rubrics. Rubrics provide scoring criteria to be used in evaluating student responses. In their operational state, rubrics are used by trained raters to evaluate test-taker responses in a standardized and consistent way, applying common criteria. Draft rubrics also play a central role in the task design and development process. Rubrics are developed in parallel with Item Specifications, and it is essential to have a clear correspondence among the contents of the task directions, the task characteristics (as

defined in the Item Specifications), and the rubrics, which define the expectations for students in responding to the tasks.

The draft ELPA21 rubrics were developed during the same time period as the Item Specifications. A rubric for each task type was created by describing expected student responses, that is, what students at varying levels of proficiency would be expected to say or write in response to an item. Because the Enhanced Assessment Grant timeline did not allow for prototyping or pilot testing of CR items, sample student responses were drafted by the ETS Team.

In drafting the rubrics, careful attention was paid to the ELP Standards and the PLDs. Because the PLDs contain expected learning progressions, their wording was frequently incorporated directly into the rubrics. The PLDs also provided guidance on what should and should not be considered in evaluating student responses. For example, there are several ways to assess proficiency in the domain of writing; however, the standards and PLDs serve to limit assessment to a subset of areas in that domain. Mechanics such as spelling and punctuation are not included in the ELP Standards and, thus, are not included as evaluation criteria in the rubrics. These aspects of writing are assessed by content-area English language arts assessments.

A total of 42 holistic rubrics were developed for ELPA21, 28 for speaking and 14 for writing. Depending on the complexity of expected responses to a task type, rubrics contain points ranging from 0-2, 0-3, 0-4, or 0-5. All rubrics contain an overarching descriptor for each score point (e.g., The response succeeds in meeting the communicative demands of the task.) followed by a bulleted list detailing characteristics of a typical response at that score point. All rubrics include 0 as a possible score, to be used for responses that do not attempt to meet the communicative demands of the task, contain no English, or do not address the prompt. While rubrics for a task type that is used at multiple grades or grade bands are similar, a different rubric was created to reflect the criteria included in the PLDs for each grade or grade band. The point range for each task was established first by creating expected responses for a sample item for each task type. Additional expected responses for each score point on a rubric were also created. The rubrics were developed so that the range of responses predicted by generating expected responses could be scored.

IAD TMT reviewers were given the opportunity to suggest edits to the initial versions of the draft rubrics for all grades early in the development of the item specifications. Based on reviewer comments, changes were made to the rubrics for one grade band, Grade 6-8. Updated versions of the Grade 6-8 rubrics were sent for a second round of review by the IAD TMT. Revisions were made to rubrics for all grade bands based on feedback from the IAD TMT's review of the Grade 6-8 rubrics. All rubrics then went through an internal review process at ETS where experts in CR scoring reviewed the rubrics for consistency, ease of use in scoring, alignment to the PLDs, and match to task type. The rubrics were updated in preparation for the content and bias panel meetings. Time did not allow for an additional round of IAD TMT review prior to the content committee meetings.

All rubrics developed under the current scope of work have been labeled as “draft.” These draft rubrics were developed through careful analysis of ELP Standards and the PLDs and in close coordination with the development of the Item Specifications. However, the rubrics have not yet been validated through use with actual student responses. Once student responses are available (i.e., once a representative sample of student responses from the field test have been received), best practice dictates that rubrics should be revised before use in live scoring of the field test responses. Such a validation effort should evaluate and refine the rubrics in response to questions such as:

- Do student responses in practice spread appropriately over the number of score points in the draft rubrics? (At this point, rubrics can be revised to have fewer or more score points as needed to appropriately distinguish among classes of student responses.)
- Do student responses include examples of the criteria listed for each score point? (If not, criteria for which no examples elicited should be removed, and new criteria should be added to reflect typical patterns of student responses.)
- Are criteria clear for each score band clear enough for raters to apply consistently? (If not, criteria should be refined and clarified as needed, with particular attention to distinctions between score bands.)

After the rubrics have been through such a validation process, they can be used to locate benchmark and range-finding sample responses for each score band and task type and to train raters for live scoring of the field test (and subsequent operational) responses.

### *3.2.4 Draft Scoring Rules for Technology-Enhanced Items*

As noted in Section 1, the provision of a range of innovative task types, including task types defined as TE, was a priority for ELPA21 (and was required by the Enhanced Assessment Grant that provided funding for ELPA21). For the purposes of ELPA21, TE items are defined as those computer-delivered items that include specialized interactions in the student response format or in the use of response data. While a range of innovative items are included in the ELPA21 item pool (including technology enabled items, those that use digital media as all or part of the stimulus), particular emphasis in ELPA21 was placed on TE items.

For ELPA21 the following TE item types (examples of which are included in the Item Specifications) have been included in the field test item pool:

- Drag and drop
- Dropdown
- Click on sentence
- Hot spot
- Audio response

Responding to a TE item requires a specialized interaction that can be more complex than responding to SR (multiple choice) or text-entry (keyboarding/typing) items. Because TE items can include more complex interactions, responses may be richer and may show evidence of

different levels of performance. Thus, some TE items are expected to be eligible for partial credit scoring. For ELPA21 TE items, the guiding principle was to assign partial credit scoring rules only when the expected response might distinguish multiple levels within the standard. That is, partial credit scoring rules were assigned only when student responses to an item might provide evidence of different levels of proficiency.

ETS test developers reviewed each TE item in the pool to determine which items might be eligible for partial credit scoring following the principle outlined above. Test developers then made judgments as to whether the student interactions addressed two or more PLDs of the aligned standard(s) and would, therefore, provide distinct information regarding the student's English language proficiency. These preliminary scoring rules for awarding partial credit were then reviewed by the IAD TMT. The IAD TMT reviewed and provided input on the proposed scoring rules, and draft scoring rules, reflecting the TMT's input, were then entered into the metadata for each affected item in IBIS (as described in more detail in Section 3.2.4).

It is important to note that the scoring rules for the TE items are preliminary and will need to be validated with actual student data once the field test has been administered. There is a plan in place to conduct such a scoring rules validation for the partial credit TE items using a modified range-finding approach. This validation process will analyze the psychometric results of the item analysis of TE items and will confirm existing rules or recommend modification or addition of rules as appropriate.

### *3.2.5 Item Banking and Metadata*

For an innovative, computer-delivered assessment system such as ELPA21, which features a considerable proportion of innovative task types, the approach to item banking and to metadata are particularly important parts of the item design and development effort.

#### *3.2.5.A Item Banking*

The item bank for the ELPA21 item development effort was IBIS, ETS's scalable system that is capable of supporting the end-to-end item banking process from item acquisition and review, through form assembly, test delivery, and statistical review.

For ELPA21, IBIS was used as the item bank of record during the item development and review process. It was used for creation, storage, and tracking of all items and metadata including initial authoring, internal ETS reviews, reviews by TMTs, reviews by state educators, and for export of XML item content and metadata to the ELPA21 field test delivery vendor.

As noted in Section 1, portability and interoperability were important requirements for the ELPA21 item pool as item design and development, field testing, and operational administration could all potentially be performed by different organizations contracted by the ELPA21 Consortium. Portability is a strength of the IBIS system, as IBIS item exports conform to the

industry-standard APIP v1.0. This compliance is certified by the IMS Global Learning Consortium, which publishes the standard.

APIP provides explicit data models for structuring elements of item content (passages, stems, response options, etc.) so that they can be understood and interpreted across a range of systems and platforms. APIP does not, however, encode display or other rendering specifications for the content that it structures. As IBIS is an item banking system, the structure of the content does not encode stylistic specifications for the art or audio files (e.g., size of files, type of files), how the item should be displayed (e.g., how a passage is displayed with its item[s]), or certain aspects of editorial style (e.g., use of bold font in item stems).

Additionally, although IBIS provides a mechanism to preview item rendering using an ETS delivery system, this does not necessarily reflect how an item will ultimately be delivered using a different system. In the ELPA21 review process, this was something of a limitation, as reviewers, in some cases, would have been able to provide more comprehensive reviews had they been able to see items with the exact functionality as they would appear in the assessment's actual delivery system.

### 3.2.5.B Metadata

In an innovative, next-generation assessment system such as ELPA21, the metadata accompanying each item plays an important role in the assessment design as it will be essential to later evaluation of item performance. An extensive number of metadata fields, coded to each item in the pool, is needed for a variety of purposes including pool inventory, field test assembly and evaluation, and future research studies.

Metadata requirements for ELPA21 were established by ETS in collaboration with the IAD TMT. The metadata fields to which each ELPA21 item are coded include:

- A unique identifier for each item and for each passage or stimulus
- Associated grade or grade band (K, 1, 2-3, 4-5, 6-8, 9-12)
- Modality (interactive, productive, or receptive)
- Item type (based on response format: SR, TE, short CR, extended CR)
- Task type and sub-type (as defined in the Item Specifications)
- Academic content area correspondence (ELA, math, science)
- Domain (Listening, Reading, Speaking, Writing)
- ELP Standard(s) assessed (1-10)
- Sub-claim(s) assessed
- PLDs
- CCSS/NGSS practice(s) assessed
- Accessibility concerns
- Accessibility features
- Experimental information
- Relationships to other items in the pool (including “parent” of twin items or “do not include with” for experimental items)

- Key
- Text complexity (for reading passages, grades 2-12)
- Item writer (allowing identification of items originating from educators from the consortium states or from ETS)

Once the metadata fields and available values for each were confirmed by the IAD TMT, decisions were made regarding the contexts in which metadata would be made available. Some metadata were chosen for inclusion on item cards while others were to be made available via separate reports to be run from the item banking system. Additionally, schema for coding the metadata in the IBIS system was established.

At this point, sample item cards (for use by the content and bias review committees) were generated and revised based on input from the IAD TMT. These item cards, and the metadata on them, were a key point of review and discussion at the content and bias review committee meetings, with several revisions to metadata coding made based on input from the educators on those committees.

Before ELPA21 items were entered into the IBIS system, ETS developed a range of process documents to guide the work of item entry and review. These included item writing templates (to ensure that all items, as drafted, contained required content elements and metadata); IBIS templates for entering APIP-compliant XML; metadata schema defined in IBIS (enabling metadata to be selected from pre-defined valid values via drop-down menus, removing the potential for mistyping); trainings for staff performing item entry; and quality control procedures for item entry, approval, and export. The process of exporting the item pool is described in Section 3.3.6.

Although robust process documents were created, the accelerated timelines of the ELPA21 project meant that some conceptual decisions related to item features and metadata categories were still being clarified or revised during the item creation and review process. As a result, a relatively extensive review and clean-up process was needed after item content had been finalized and before the item pool would be ready for export.

### *3.2.6 Editorial Style Guide and Graphics Style Guide*

An Editorial Style Guide and a Graphics Style Guide were developed to help ensure that presentation of content for each grade or grade band was consistent, effective, and grade appropriate. Development of these style guides occurred roughly in parallel with the development of the Item Specifications.

For both the Editorial Style Guide and the Graphics Style Guide, ETS began by drawing on existing documents; at the direction of the TMTs, ETS made substantial use of process documents created by the Smarter Balanced Assessment Consortium, with enhancements and refinements to meet the needs of ELPA21. (For example, the ELPA21 Style Guide includes decisions that address content presentation for students in Kindergarten and Grade 1.) Initial

drafts of the ELPA21 style guides were reviewed by the TMTs, and changes and refinements were made based on TMT input.

#### *3.2.6.A Editorial Style Guide*

The purpose of the Editorial Style Guide is to establish a clear and grade-appropriate representation of language for each grade or grade band. It includes both general style considerations related to computer-based testing (including best practices related to presentation of content panes and scrolling) as well as guidance on editorial style, word usage, punctuation, and writing of directions, item stems, and item options.

A topic of extensive discussion related to the Editorial Style Guide was the identification of an appropriate delivery font for each grade or grade band. A range of experts on the AAA TMT, within ETS, and others were consulted to identify fonts that would be easy to read on screen and would be age appropriate, particularly with regard to best letter forms for emerging readers. In addition, ELPA21 placed value on identifying a font that was available without charge to avoid any future financial obligations on behalf of consortium member states. In the end, the decision was made to use 18-point Comic Sans for Kindergarten and Grade 1, and to use Verdana for Grades 2-3 and above (in sizes ranging from 18 point for Grades 2-3 to 14 point for Grades 6-8 and 9-12). These decisions are documented in an appendix to the Editorial Style Guide.

#### *3.2.6.B Graphics Style Guide*

Art, illustrations, and other graphic elements play a very prominent role in ELPA21. Because ELPA21 is an assessment of English language proficiency, the ELPA21 task types generally rely heavily on graphics to communicate with students and provide stimuli for them to respond to in English, particularly for students in Kindergarten and Grade 1 who are expected to be at a fairly early stage of literacy development. The Graphics Style Guide includes a conceptual explanation of how ELPA21 graphics are to be developed in accordance with principles of Universal Design; technical guidelines for issues such as art formats, color palette (chosen with attention to accessible colors for students who are color blind), and file delivery; as well as detailed guidelines for and examples of illustrations depicting students, teachers, and common classroom objects. The example illustrations, as well as initial samples of illustrations produced for field test items, were a particular focus of discussion and input during TMT reviews, resulting in an updated series of guidelines and examples for producing art that was consistent and age-appropriate for each grade or grade band.

Because of schedule requirements, one limitation of the Graphics Style Guide is that the document, and the ELPA21 art produced based on it, were created before the field test delivery vendor had been identified and, therefore, before details regarding the technical requirements, capabilities, and limitations of the field test delivery system were known. As a result, although all ELPA21 graphics were produced according to reasonable and standard practices, specific requirements needed for construction of the art based on the capabilities of the field test delivery vendor were not available at the time the art was developed. Consequently, the items and their graphic elements could not be vetted in advance to evaluate their compatibility with the delivery system of the field test delivery vendor.

### 3.2.7 *Sample Items*

As mentioned in section 3.2.1, a selection of sample items was created during the development of the Item Specifications to illustrate the ELPA21 task types. Because there is overlap in task type across grades, it was not necessary to develop a sample item for each task for each grade or grade band. Instead, sample items were created for all task types used in Kindergarten, Grade 1, and Grades 6-8. The initial sample items were reviewed by the IAD TMT in a paper format at various stages of the development of the Item Specifications. Feedback on the content and appropriateness of the sample items was received and acted on at numerous stages of the Item Specifications review. After the first round of IAD TMT review, additional sample items were developed to reflect those task types in Grades 2-3, Grades 4-5, and Grades 9-12 not found in the initial set of sample items, and the creation of functional computer-based versions for a subset of the sample item collection began. Table 13, in Appendix A, provides details on the number of sample items produced.

The ELPA21 sample items were used for a variety of purposes throughout the development and export of the ELPA21 item pool. Initially, the sample items were used as exemplars during item writer training. In combination with the detailed descriptions for the required elements of each task, the sample items provided item writers with concrete models to follow.

Once computer-based versions of the sample items were created, these samples served as an advance set of items that would move through the item production and export process from start to finish, allowing, in effect, a “dry run” of downstream processes. These were the first ELPA21 items to be entered into the ETS item bank. As these items were entered, training documentation for future item entry was written. These items went through the various stages of internal review, including review for accessibility and APIP. In addition, professional recordings were made for audio components of the sample items. When the field test delivery vendor for the ELPA21 was announced, a subset of the sample items was prepared for export to the vendor. This was an important step in beginning to plan for the item transfer process that would take place when the entire pool of items was exported from ETS to the field test delivery vendor.

In addition, when a small-scale cognitive laboratory study was proposed for ELPA21, the question arose as to what an appropriate source of test items would be. Because the field test pool needed to be kept secure, it was not desirable to use intended operational items. The sample items were a viable option, and the field test delivery vendor was able to prepare the sample items for use in the cognitive laboratory study.

Finally, as ELPA21 prepared for field testing, a need arose for a selection of sample items for teachers and students to become familiar with the ELPA21 task types. Again, the sample items were available for this purpose. The sample items were reviewed and a subset were selected for use in an interactive demo to be made available for consortium teachers and students.

### 3.3 Item Development

#### 3.3.1 Item Development Plan

The overall goal of the ELPA21 item development effort was to create enough field-test-ready items to support subsequent development of operational forms for both a screener and a summative assessment. The Item Development Plan was created as the guiding document for developing the ELPA21 field-test-ready items.

The plan described three major item development goals:

1. **Produce High-Quality Items.** Ensure that items:
  - Align with ELP Standards;
  - Assess an appropriate range of proficiency as described by the PLDs;
  - Follow the ELPA21 Item Specifications;
  - Satisfy the agreed upon number and distribution of items according to grade band, domain, and item type.
2. **Effective Use of Resources.** Effectively leverage the knowledge, skills, and abilities of those involved in the item development process, including educators from the consortium states, members of the IAD and AAA TMTs, ETS staff, and contracted item writers.
3. **Timely Delivery of Items.** Make items available in IBIS for client review and for content and bias review according to the project schedule.

The plan summarized the key tools, activities, and processes that were carried out to support the production of high-quality ELPA21 field-test-ready items. The tools that item writers and content reviewers used to guide the development of high-quality items were:

- ELP Standards
- ELPA21 Item Specifications
- Sample items
- Templates for entering item metadata and content
- Checklists for reviewing items

The ELP Standards and the ELPA21 Item Specifications were the primary references for developing and reviewing items. All item writers received the same general training regarding ELPA21 and principles for item development. Item writers then received specific training regarding the task types for which they were to develop items. During the specific training, item writers became familiar with relevant portions of the ELPA21 Item Specifications as well as relevant sample items. Standardized item templates required item writers to enter information that was needed for item entry in IBIS. A detailed checklist for reviewing items was developed and used by both item writers and content reviewers. Uniform general training, focused training on specific item types as described in the ELPA21 Item Specifications, and training to enter

items using standardized templates all promoted efficient development of items that were aligned to ELP Standards.

To promote development of a pool of items covering an appropriate range of domains and item types across all grades and grade bands, the Item Development Plan includes tables of the number and distribution of items to be developed. According to the plan, approximately 449 items would be needed to develop a single operational form across all grades or grade bands. Since some items are likely to be rejected after field testing, it was estimated that at least 2,024 items would need to be field tested to yield a robust pool of items for operational form development. Similarly, because of expected attrition during content and bias committee reviews, it was estimated that 2,454 items would need to be developed to ensure that at least 2,024 would be approved for field testing. The Item Development Plan also detailed the requirement for review of items in four batches by TMT members prior to the content and bias review committee meetings.

The Item Development Plan also promoted effective use of resources. It provided information regarding the teams who were responsible for securing passages, drafting items, and reviewing items, as well as individual team member roles. The four teams of item writers who participated in the development of ELPA21 items as described in the Item Development Plan are described below.

1. ETS Core Team. The nine assessment specialists of this team developed the ELPA21 Item Specifications, sample items, and item writer training materials. Six assessment specialists from the ETS Core Team also served as ETS content leads (one content lead per grade or grade band), and, in addition to specification, item, and training materials development, they coordinated item writing and review assignments for their grade or grade band. The ETS Core Team also led item writer training and committee reviews and refined input from all parties to ensure the quality of the item pool.
2. ETS Assessment Specialists. An extended team of experienced ETS assessment specialists reviewed and revised items to ensure that they assessed appropriate ELP Standards and conformed to Item Specifications.
3. Outside Item Writers. External contract item writers with a record of developing quality items for other ETS ELL assessments also drafted items and developed listening stimuli.
4. Educator Item Writers. Educators from the consortium states drafted a variety of items. It was essential for educators from the ELPA21 consortium states to provide as much meaningful input as possible in the item development process. Educators are familiar with the ELP Standards, the needs of ELLs within their states, and the ways in which ELLs are likely to interpret the language of items. To gain a high level of educator input, reading passages and listening stimuli were developed before the educator item writer training. Educators were then asked to draft items specifically for these passages and stimuli that assessed appropriate content and skills as described by the ELP Standards. Educators were also asked to brainstorm topics for speaking and writing CR items. The educator item writer training was an ideal venue for brainstorming and vetting topics because educators had the opportunity to share whether specific topics were

accessible to and appropriate for students within their states. Due to the fact that reading passages needed to be reviewed prior to item development, educator item writers did not have the opportunity to write reading passages. Some educators had the opportunity to write listening stimuli during assignments after the educator item writer training.

The last goal of the Item Development Plan was to outline a strategy for delivering the draft pool of items for the content and bias committee reviews in a timely manner. The plan described a scheduled series of activities from the first item writer training on May 3–4, 2014, and the educator item writer training on May 28–29, 2014, to the content and bias committee reviews of August 18–22, 2014. In order to meet the aggressive timeline of developing and delivering the items for committee reviews in approximately four months, the schedule included a number of efficient and cost effective measures that would allow ETS to draw on the expertise of outside item writers and the extended team of ETS assessment specialists. The schedule included a special Spring Item Writing Institute, which employed six outside item writers for two months to develop and review items. The schedule also included intensive work marathons (i.e., on-site item writing and review sessions performed by the extended team of ETS assessment specialists during specified blocks of time).

The Item Development Plan, which was drafted and revised according to client feedback, provided a detailed summary of key activities and processes to support the high quality and timely delivery of ELPA21 field-test-ready items.

### *3.3.2 Reading Passage Acquisition and Review*

#### *3.3.2.A Reading Passage Acquisition*

One key goal of the ELP Standards, like the CCSS to which they correspond, is that students will be able to read complex, grade-level texts independently. In light of this goal, the Item Specifications call for ELPA21 reading items to be based on literary and informational reading passages at a range of complexity levels. ELPA21 also had a goal of including authentic and original materials in the test as a reflection of the texts that students will need to interact with in the real world. To help meet that goal, a target was set to acquire 20 percent of reading passages for the field test pool from authentic (i.e., previously published) sources. Public domain passages were located to fulfill this requirement. In addition, passages were commissioned from passage writers with experience in writing K-12 texts. Passage writers were given guidelines for length, topics, and text complexity. Once passages were received by ETS, they were reviewed for appropriateness for the intended task type and grade or grade band.

Passages judged by the ETS Core Team to be potentially suitable for use were then sent for two levels of review: an initial screening by state representatives and then a more intensive review, conducted via WebEx, with educators from the consortium states.

### 3.3.2.B State Reading Passage Screening Process

The high-level screening by the state representatives served to ensure that any passages containing material that was potentially objectionable or otherwise inappropriate for one or more of the consortium states would be removed from consideration and no further effort would be placed into developing them.

The initial pool of passages acquired by ETS was delivered to state representatives, who organized and managed the screening process independently for each state. For example, Oregon participated in this screening with a team of EL teachers, math teachers or specialists, and science specialists (all at the K-16 level). Some Oregon Department of Education (state-level) specialists participated. Feedback from Oregon was submitted with extensive notes via the surveys that were provided by grade band. Oregon used majority rules for the comments and recommendations although, if there was an outlier comment that might be of interest, it was noted and sent to ETS. The state representatives were asked to respond the question: “Is this passage free of potential bias and topics that may be sensitive or inappropriate?” If the answer was “no,” they were asked to provide a short explanation.

A total of 247 passages were reviewed during two rounds of state screening. During the initial state screening, 219 passages were reviewed before the educator passage review. For Grade 1, a small number of passages received too late for the initial state screening process were included in the educator passage review, and then routed for post-hoc state screening (at which all were deemed acceptable). In total, of the 247 passages that were screened by state representatives, 18 were rejected as shown in Table 6 (in the following section). ETS removed the rejected passages from the pool, and made revisions to some passages as suggested by the state representatives.

### 3.3.2.C Educator Passage Review WebEx Meetings

The educator passage review meetings, which constituted the first of the three major opportunities for educators from the consortium states to have substantive input to the ELPA21 item pool, were considerably more intensive than the state screening process.

Educators from the consortium states were recruited by ELPA21 state representatives. (See Appendix B for a list of educators who participated.) The educators were organized by ETS into three panels: K, Grade 1, and Grades 2-3; Grades 4-5 and Grades 6-8; and Grades 9-12. Because passages are longer at the higher grades, the panels were organized so that each set of educators would have a similar volume of materials to review. The panel meetings were held via WebEx over the course of three consecutive days and began with training on the following topics:

- An overview of ELPA21 (including the ELP Standards, the ELPA21 mission, and the item development process);
- The reading task types selected for inclusion on ELPA21, including the specifications for informational and literary passages at each grade or grade band;
- The review criteria to be used in the passage reviews;
- The process for reviewing and commenting on passages.

The review criteria asked the educators to consider three prompts:

1. Is the topic appropriate for ELPA21 students in this grade or grade band?
2. Is the passage free from issues of bias or sensitivity that would offend or disadvantage any identified group?
3. Please rate the passage's complexity and general level of challenge for EL students relative to the identified grade/band.

Over the course of the WebEx meetings, educators were given time to read the passages and consider the questions before coming together for discussion. ETS Core Team members, acting as facilitators, used the polling feature of the WebEx interface to ensure that all educators on each panel participated. In many cases, the educators worked together to make improvements and refinements to the passages.

Once discussion was complete, the educators provided their final rating of each passage, placing it into one of three categories:

- accepted as written (i.e., with no changes);
- accept with revisions;
- not acceptable (i.e., reject).

Based on the results of the initial poll, the facilitators then led discussion as appropriate to reach a majority group consensus of educators. Educators were able to submit comments via WebEx chat function and could speak via phone. The facilitators were tasked with asking questions that helped to clarify educators' opinions, asking educators to give reasons to support the status they selected, etc. All suggestions for revisions were documented. Some changes were made in real time on the screen so all educators could approve the changes. Passages that were deemed not acceptable were removed from the passage pool and were not developed further.

The results of the educator passage review, along with the preceding state screening, are shown in Table 6.

**Table 6: Reading Passages: Results of State Screening and Educator Review**

Grade or Grade Band	State Screening		Educator Passage Review			
	# screened	# rejected	# reviewed	# approved as written	# approved with revisions	# not acceptable
<b>K</b>	36	2	34	0	34	0
<b>1</b>	53	6	47	15	31	1
<b>2-3</b>	49	0	49	1	48	0
<b>4-5</b>	38	5	33	0	33	0
<b>6-8</b>	34	1	33	0	32	1
<b>9-12</b>	37	4	33	27	4	2
<b>Total</b>	<b>247</b>	<b>18</b>	<b>229</b>	<b>43</b>	<b>182</b>	<b>4</b>

### 3.3.3 Item Writing

#### 3.3.3.A Item Writer Recruitment

ELPA21 required that 25 percent of items originate from educators in consortium states who were identified by the ELPA21 council. To supplement the items written by the educators, ETS internal item writers and external contract item writers produced items as described in the Item Development Plan.

The ETS internal item writers were full-time assessment specialists in the English Language Learning division with either Bachelor's or Master's degrees in English language education, linguistics, or related fields and with experience teaching English as a Second Language. The nine assessment specialists on the ETS Core Team devoted a majority of their time to ELPA21 item development throughout the duration of the project. In addition to leading development of the Item Specifications, this core team of ETS assessment specialists was responsible for following appropriate procedures to ensure all items were developed to align to the ELP Standards.

The educators from the consortium states were recruited and selected by ELPA21 state representatives. ELPA21 implemented an application and selection process to recruit a diverse and representative group of educators with classroom experience with ELs. (See Appendix B for a list of educators who participated.)

The final set of item writers was a group of six external contract item writers who were recruited by ETS. All six item writers were selected on the basis of their strong performance as item writers for other EL assessment programs.

#### 3.3.3.B Development of Item Writer Training Materials

In preparation for item writer training, training materials were developed by the ETS Core Team, reviewed by the IAD TMT, and then revised in response to the IAD TMT's feedback. The training materials consisted of a PowerPoint presentation called *ELPA21 General Item Writer Training*, a handout about ETS Fairness Guidelines (adapted from the *ETS Guidelines for Fairness Review of Assessments*), a Checklist for Reviewing Test Items, Item Specifications, sample items, and item templates.

The *ELPA21 General Item Writer Training* presentation included essential information for all item writers. Topics that were covered included:

- Overview of ELPA21 and its mission;
- Item development process;
- ECD;
- Seven elements of Universal Design;
- Good item writing practices based on Universal Design;
- APIP;
- Fairness Guidelines;
- Overview of ELP Standards;

- Overview of English language arts (ELA), math, and science practices;
- Overview of PLDs;
- Overview of the checklist for reviewing items;
- Overview of the grade-level training with teams in separate rooms;
- Process for submitting items after the training.

The presentation included several examples of poorly constructed items for trainees to review and discuss so that they had an opportunity to apply new concepts.

Two other documents were designed for review during the general presentation: the Fairness Guidelines handout and the Checklist for Reviewing Test Items. The Fairness Guidelines handout described the crucial need for avoiding issues of sensitivity or possible bias in ELPA21 items and provided a list of the types of topics to be avoided to allow students to demonstrate their full English language abilities. The Checklist for Reviewing Test Items provided detailed guidelines for developing quality reading passages, listening stimuli, graphs/charts, questions, and options (for multiple choice items). The checklist was designed as a tool for item writers to critique their own work before submission.

The remainder of the training materials was prepared for grade-level training with teams in separate rooms. For each task type, assessment specialists provided the Item Specifications, sample items, and item templates. The Item Specifications described all of the features of the item type, including the types of passages, stimuli, artwork, questions, directions, and response types to be developed. The sample items provided item writers with good examples of the item types. The item templates provided a structure for item writers to use in entering and submitting the content of their own original items.

The item writer training materials were used twice: first, during the training of internal and contract item writers on May 3–4, 2014, and next, during the training of consortium educators from May 28–29, 2014. A description of these two training events follows.

### *3.3.3.C Training of ETS Internal and External Contract Item Writers*

The training of ETS internal item writers and external contract item writers took place over a two-day period from May 3–4, 2014, in Princeton, New Jersey. The training materials were early versions of the materials that were used to train the consortium educators and, thus, served as a trial run of the subsequent training for educators from consortium states.

As planned, the training began with a general PowerPoint presentation called *ELPA21 General Item Writer Training* that described key principles of guiding the development of high-quality items. The presentation covered key aspects of ECD, Universal Design, Fairness Guidelines, and ELP Standards. After the general training, participants divided into groups and received training from the ETS content leads regarding the specific types of items to be developed and the particular ELP Standards that they were designed to assess. Item writers developed items during the two-day event and received feedback from ETS content leads. Item writers also responded to each other's draft items to determine whether they elicited the type of language as described in the aligned ELP Standards. ETS content leads and item writers also spent time brainstorming

topics for task types that required listening stimuli and then selected acceptable topics to develop after the two days of training. During the two-month period after the initial training, item writers completed assignments to develop a variety of task types, with ETS staff providing ongoing feedback.

#### *3.3.3.D Training of Educator Item Writers*

The educator item writer training took place over a two-day period from May 28–29, 2014, in Dallas, Texas. A total of 52 educators attended the item writer training from the consortium states of Arkansas, Iowa, Kansas, Nebraska, Ohio, Oregon, Washington, and West Virginia. Two educators from the National Center on Educational Outcomes were from the state of Minnesota. (See Appendix B for a list of educator participants.)

The training materials that were used during the May 3–4 training at ETS were updated in response to feedback and then used to train the consortium educators. As with the May 3–4 training, the consortium educator training consisted of two types of training: a general training session for all participants and grade-level training that was provided in small groups. The ETS Core Team who worked on the development of the ELPA21 Item Specifications and sample items served as the trainers.

The general training session covered key aspects of ECD, Universal Design, Fairness Guidelines, and ELP Standards. After the general training session, educators met in small groups according to grade or grade band to develop items. Much of the item development effort focused on items to accompany reading passages. Educators reviewed Item Specifications and sample items to become familiar with the types of content that items needed to assess as well as the number of each type of item that needed to be developed for each passage. Educators also received training and completed assignments to develop listening items at the grades of Kindergarten, 1, 2-3, and 4-5. The ETS Core Team members in each room answered questions as they arose and reviewed items as they were submitted. Educators also spent time brainstorming and developing ideas for speaking and writing CR items. This was an ideal venue for developing CR prompts because educators were able to brainstorm a variety of ideas, discuss the ideas, and keep those that were accessible and appropriate for the students at that grade or grade band across the consortium states.

Since some schools were still in session, it was optional for consortium educators to accept assignments after the training event. A number of consortium educators completed assignments and submitted additional items after the two-day training.

The training was highly successful yielding a total of 1,290 draft items across all grade bands. Although there was naturally some attrition of educator-originated items as they moved through the various stages of the review process, 920 (or 37 percent) of the items in the field test pool as delivered originated from educators, comfortably exceeding the target of 25 percent. Educators also succeeded at producing a wide variety of items. Educators at five of the six grades or grade bands developed items within each of the four domains of listening, reading, speaking, and writing. At Grades 9–12, teachers developed items from three domains; they did not develop

speaking items because of the high numbers that were developed at ETS prior to the educator item writer training.

Table 7 provides details on the number of items developed by ETS internal and external item writers and consortium educators that were eventually approved for field testing.

**Table 7: Origination of Intended Operational Items Approved for Field Testing**

Domain	Item Writer	K	1	2-3	4-5	6-8	9-12	Grand Total	Percentage from Educators
Listening	ETS	154	79	102	80	118	117	650	
	Educator	15	44	15	45			119	15%
<b>Listening Total</b>		169	123	117	125	118	117	769	
Reading	ETS	88	87	60	63	82	58	438	
	Educator	57	69	75	62	30	97	390	47%
<b>Reading Total</b>		145	156	135	125	112	155	828	
Speaking	ETS	122	13	61	60	34	54	344	
	Educator	1	62	4	25	19		111	24%
<b>Speaking Total</b>		123	75	65	85	53	54	455	
Writing	ETS		5	23	36	27	26	117	
	Educator	91	70	70	36	12	21	300	72%
<b>Writing Total</b>		91	75	93	72	39	47	417	
<b>Grand Total</b>		528	429	410	407	322	373	2469	
<b>Total Items Originating from ETS</b>								1549	
<b>Total Items Originating from Educators</b>								920	37%

Consortium educators provided feedback in a survey at the end of the two-day event. Overall feedback from the consortium educators was strongly positive, indicating that the training succeeded in providing a meaningful professional development experience to participating educators. Of the 49 educators who provided an overall rating of the event, 47 said they would describe the event as “good” or “excellent” while only 2 described it as “fair” and none as “poor.” Of the 53 educators who answered the question of whether they would participate in this event again, 52 said “yes” and only 1 said “no.”

The following quotations from educators who participated are representative of the feedback received.

“This was an amazing experience. It was so great to have the opportunity to be a part of the test writing process, as it directly impacts the students we teach. There was so much work that had been done prior to our training days. This allowed us to be very productive! In addition, throughout our time in Dallas, we had the opportunity to connect & learn from other ELL educators from across the country.”

“I really appreciated your efforts in including educators in the assessment item writing. You not only include us on something that may determine and label teachers as effective or not, but take the time of teaching us and sharing your expertise with us. We as teachers

create tests and regularly offer formative assessments but are not taught the specifics on how to do this. Thank you for including us!”

“The workshops allowed for me to learn more about the new assessment. Being a part of the training helped me begin to make connections that I will use to guide my staff in their understanding of both the assessment and its connection to both standards and content. I’m most excited by the emphasis the assessment places on academic language. The ELPA21 team has really taken time to understand language in relation to ELLs, and that is evident through the facilitator’s knowledge and the information shared/created.”

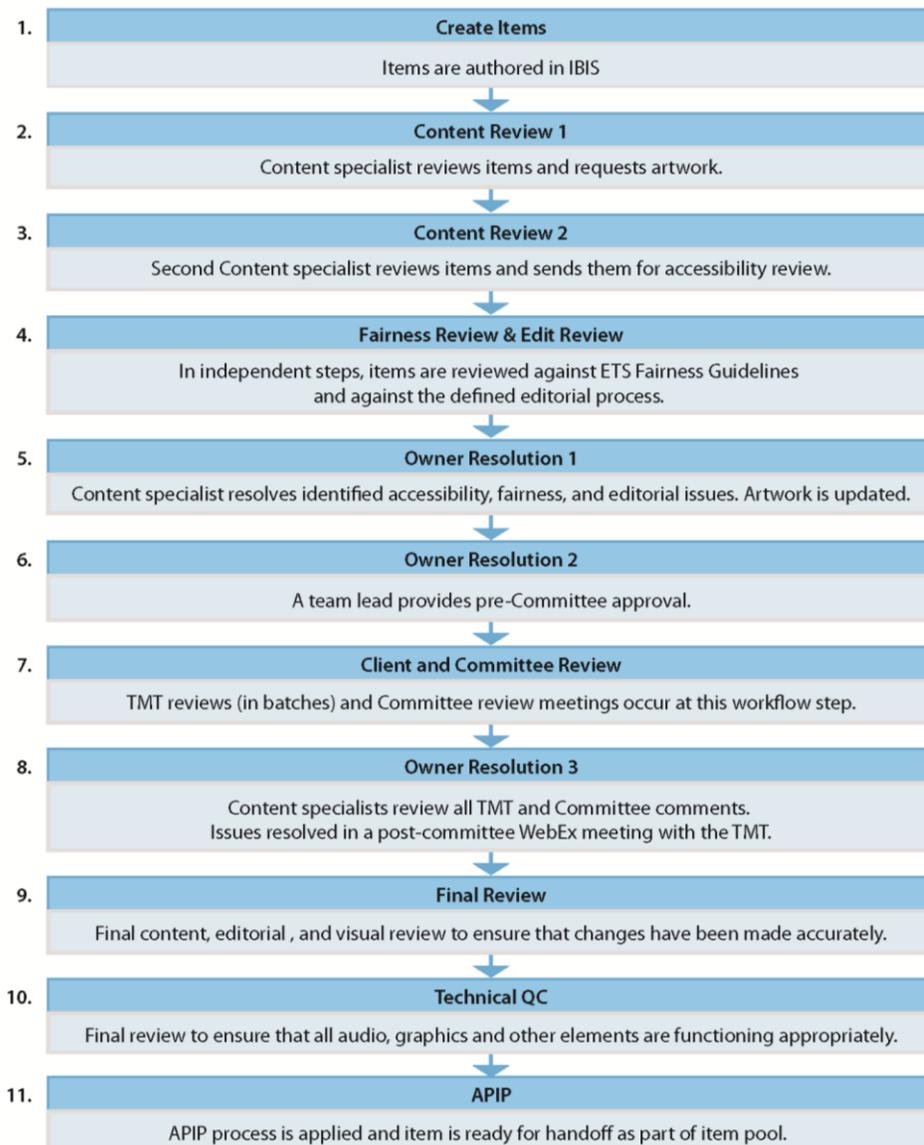
The comments provided above and the overall positive nature of the feedback indicated that the item writer training event succeeded not only in generating draft items for ELPA21, but also in providing meaningful professional development to participating educators regarding the ELPA21 test development process, positioning them to be “ambassadors” for ELPA21 when they return to their schools and districts.

3.3.4 Item Review

3.3.4.A Internal ETS Reviews

As discussed in Section 3.2.5, ETS uses IBIS to manage workflow, ensuring that all items move through every planned review step. Figure 1 shows the ETS item development process workflow for ELPA21.

**Figure 1: ETS Item Development Process for ELPA21**



Note: “Item Author” in this figure refers to entry into IBIS, not to writing the content of an item.

After items had been authored (i.e., entered) in IBIS, they went through a series of rigorous content reviews to ensure that each item was of high quality. During the content review, assessment specialists reviewed each item to ensure that it contained all assets as described in the ELPA21 Item Specifications. Assessment specialists also reviewed and revised items according to shared criteria. The following are some of the questions that assessment specialists considered as they reviewed items:

- ✓ Does the item test knowledge or a skill or an ability that is called for by the ELPA21 Item Specifications?
- ✓ Does the item align to the identified ELP Standard(s)?
- ✓ Is the item within the appropriate range of difficulty as described by the PLDs?
- ✓ If the passage was selected from a larger text, does the passage make sense on its own?
- ✓ Is the language structure and complexity grade-appropriate?
- ✓ Is the vocabulary grade-appropriate?
- ✓ Does the listening stimulus contain appropriate content for the purpose of the test with the intended test population?
- ✓ Is the listening stimulus realistic? Is the language in an appropriate register for the situation? Is it reflective of spoken language?
- ✓ Are any graphs or charts correctly and clearly labeled?
- ✓ Do any graphs or charts provide all the information needed to answer the questions?
- ✓ In the stem, is a clearly defined problem or task posed for the student?
- ✓ Can the stem be worded more clearly or concisely?
- ✓ Are options reasonably parallel in structure and complexity?
- ✓ Do options fit logically and grammatically with the stem?
- ✓ Which option do you think is intended to be the correct response? Does your choice agree with the intended key?
- ✓ Is there a better key for the stem than has been stated among the options?
- ✓ For constructed response items: Will the prompt elicit a full range of responses as described by the rubrics?
- ✓ For TE items: Are the number and types of interactions needed to obtain a correct response appropriate?

After test materials had gone through two content reviews, ETS editors performed fairness reviews of all ELPA21 tasks, items, and passages. The goal of the ETS fairness review is to identify and remove any language or content that could be offensive or biased. Offensive topics (such as drug use or supernatural experiences) need to be avoided so that students can focus on the assessment tasks and perform to their full potential. Biased language and content (such as material that assumes knowledge of a religious holiday) need to be avoided so that the test

material is based on experiences that are accessible to all groups of students. The main questions that fairness reviewers considered as they reviewed items were:

- ✓ Does the material contain language or content that could be offensive or inappropriate for a population subgroup? Could any aspect of the material be construed as elitist, sexist, or racist? Does the material refer to an upsetting or controversial topic?
- ✓ Is any outside knowledge (e.g., procedures for making laws) or cultural knowledge (e.g., holiday customs) required to understand the material? Could an explanation be added without damaging the materials?

After fairness review, all ELPA21 tasks, items, passages, and stimuli were given a substantive editorial review. During the editorial review, each item (including text and any graphics) was checked for correctness and clarity of language, consistency of style, accessibility, and conformance to the standards set forth in the Editorial Style Guide. Lead editors provided guidance to all editorial staff to maintain consistency, document the standard application of style and grammar decisions, and increase efficacy of edit reviews.

After content, fairness, and edit reviews, the items went through a pair of “owner resolution” reviews. Within the ETS internal review process, the owner is a person who is familiar with the ELPA21 Item Specifications and is responsible for making any needed revisions to items in response to review comments. Owner resolutions of ELPA21 items were conducted by the ETS content lead or an assessment specialist who had been trained by the ETS content lead. During the first owner resolution, the owner ensured that appropriate revisions to text and artwork were made in response to all comments from the fairness and editorial reviews. During the second owner resolution, the owner ensured that all items were ready for TMT and committee reviews.

#### *3.3.4.B Pre-Committee Meeting Reviews by Task Management Teams*

One of the goals of the ELPA21 item development process was to allow an opportunity for members of the TMTs to review and comment on all items before the items were provided to the content and bias review committees. The nature of this review was to some degree determined by the short timelines available: the ETS Core Team organized the in-development items into a series of four batches and provided them to TMT members for review in IBIS over the course of several weeks prior to the content and bias review committee meetings.

Before the initial batch review, several training sessions on how to review in IBIS were organized and an informational guide was developed to provide accessible information about how to search for the appropriate items, conduct reviews, and provide feedback in IBIS. The TMT representatives who participated were able to provide a number of valuable comments that informed the content and bias review committee meetings.

The criteria for the pre-review by TMT members were the same questions that would later serve as criteria for the content and bias review committee, as shown below.

For the content review committee:

- Does the item appropriately measure the identified standard(s) and practice(s)?
- Does the item appropriately measure the identified sub-claim in a way that is appropriate to the PLD?
- Does the item have a single best key?
- Is the item presented clearly and effectively?

For the bias review committee:

- Is the item free of content that is potentially biased against or offensive to any identified group?

TMT reviewers were also given the opportunity to make broad comments that might apply to larger categories of items. Because of the short time spans (as little as a single week for some items) that was available between the end of TMT reviews and the start of content committee reviews, no immediate action was taken to implement changes based on the TMT reviewers' comments prior to the committee meetings. Instead, ETS ran reports from IBIS documenting all comments made in the TMT's pre-reviews. The ETS Core Team facilitating the content and bias review committee meetings then introduced these comments for consideration at the committee meetings; the content and bias review committees considered each comment by the TMT reviewers before deciding on the actions to be taken (i.e., accepting items as written, accepting items with revisions, or rejecting items).

#### *3.3.4.C Content Review Committee Meetings*

The goal of the content review committee meetings was to integrate input from educators across ELPA21 consortium states to ensure that items for ELPA21 field tests are aligned to ELP Standards and of the highest possible quality. ELPA21 state representatives recruited a pool of educators from across the consortium states to serve on the content review committees. A content committee of 8–12 educators was established for each grade or grade band. A total of 61 educators participated in the content committee review meetings, which took place over a five-day period from August 18–22, 2014, in Seattle, Washington. (See Appendix B for a list of educators who participated.)

Two types of training materials were prepared for the meeting: general training materials that were presented to all content review committee members at the beginning of the meetings and grade-level-specific training materials that were presented to committee members in their separate rooms. Presentations and handouts were reviewed in advance by the IAD TMT and several rounds of revisions were made based on TMT input.

The general training consisted of PowerPoint presentations that covered essential information for all content committee members. Topics that were covered included:

- Meeting agenda, purpose, and goals;

- Confidentiality and security procedures;
- Overview of ELPA21 and its mission;
- Purposes of ELPA21 assessments;
- Item development process;
- Importance of educator input;
- Scope of committee reviews;
- Goals of committee reviews;
- Role of committee members;
- Overview of ELP Standards;
- Overview of ELA, math, and science practices;
- Overview of PLDs;
- Guiding questions for content review;
- Decisions to be made by the committee;
- Overview of review process.

The general training included several example test items for discussion, which allowed participants to apply principles from the presentation. All committee members signed a nondisclosure agreement during the general training session.

After the general training, committee members went to their assigned rooms for specific training related to their grade or grade band. Group-level training was conducted by two assessment specialists per grade or grade band; these assessment specialists also served as facilitators and note takers during the five-day meeting. Assessment specialists used a checklist called Step-by-Step Guidelines for Reviews to conduct the group training so that all training sessions were conducted in a standardized manner.

It was particularly important for all committees to follow the same security protocols. All confidentiality forms were signed before committee members entered committee rooms. Item binders were checked in and out each day by the ETS assessment specialists. Educators were each assigned individual binders that were tracked throughout the week. All electronic devices were turned off during review sessions, and personal belongings, including electronic devices, were stored away from the workspace. After committee meetings were complete, all item binders were signed in for the final time and the binder contents were shredded on site.

After training, content committees reviewed items using the following criteria:

### **Guiding Questions for Content Reviews**

1. Does the item appropriately measure the identified standard(s) and practice(s)?
2. Does the item appropriately measure the identified sub-claim? If so, which PLD(s) does the item measure?
- 3a. For SR Items: Does the item have a single best key?

- 3b. For TE Items: Does the item have a correct response as described in the scoring rules?
- 3c. For CR Items: Does the question elicit information that would allow students to demonstrate their language abilities (as described by the draft rubrics)?
- 4. Is the item presented clearly and effectively?

Of particular note is that the content review committee made judgments about the alignment to the PLDs for each item; the fact that these decisions were made by a panel of educators and agreed upon through discussion helped to underscore the validity argument for this aspect of the item pool. In each case, an item that was accepted was found to have alignment to at least one ELP Standard and corresponding PLDs. Any items that did not align with ELP Standards or PLDs were rejected.

In addition to viewing printouts of items, committee members viewed select functional versions of the items as they appeared in the IBIS system, allowing committee members to see a representation of how tasks are presented and how students are to select responses. Since many of the items were TE items, it was important for committee members to have an opportunity to see the computer-based functions needed to select or enter responses.

ETS assessment specialists facilitated discussion of each item with reference to the guiding questions. The note taker read aloud comments from the TMT's pre-review for the committee's consideration as each item was reviewed. The facilitator moderated discussion regarding TMT observations or committee member observations and helped the committee reach consensus on overall decisions whether to accept an item as written, accept an item with revisions, or reject an item. When an item was accepted with revisions, the facilitator helped the committee reach consensus regarding the description of the revision. The note taker recorded the group's decisions and descriptions of any needed revisions. (See Section 3.3.4.F for a description of content review committee results.)

#### *3.3.4.D Bias Review Committee Meetings*

The goal of the bias committee reviews was to integrate input from educators across ELPA21 consortium states to ensure that items for ELPA21 field test items were free of bias and sensitivity issues and were of the highest possible quality. Consortium educators were recruited to serve on bias review committees. A bias review committee of 5–6 educators was established for each of the following grade bands: Kindergarten and Grade 1, Grades 2–3 and 4–5, Grades 6–8, and Grades 9–12. The items for these grades were grouped together in this manner because it was anticipated that reviewers at the lower grades would be able to cover more items, since item sets in the lower grades did not include lengthy reading passages. In addition, it was assumed that because the criteria for bias review were simpler, the committee would be able to review a similar volume of items in a shorter time. A total of 21 educators participated in the bias review committee meetings. The bias review committee meetings took place over a three-day

period from August 20–22, 2014 in Seattle, Washington. (See Appendix B for a list of educators who participated.)

Two types of training materials were prepared for the bias review committee meeting: general training materials that were presented to all bias review committee members at the beginning of the meetings and grade-level-specific training materials that were presented to committee members in their assigned rooms. Presentations and handouts were reviewed in advance by the IAD TMT and several rounds of revisions were made based on TMT input.

The general training consisted of PowerPoint presentations that covered essential information for all bias review committee members. The content of the general training session was similar to that of the general session for the content review committee. The overviews of general information about ELPA21, item development processes, and security measures were the same. Instead of reviewing detailed information about the ELP Standards, practices, and PLDs as the content review committee did, the bias review committee received an in-depth presentation about bias and sensitivity issues. The sample items for discussion were designed to let committee members practice using concepts about bias and sensitivity that were covered during the training.

As with the content review training, bias review committee members went to their assigned rooms for specific training related to their grade band after the general training. The ETS assessment specialists who served as facilitators and note takers used the same Step-by-Step Guidelines for Reviews to provide training regarding procedures for discussion and security protocols. All security protocols were identical to those followed by the content review committees.

After training, bias review committees reviewed items using the following criteria:

### **Guiding Questions for Bias Reviews**

1. Is the item free of content that is potentially biased against or offensive to any identified group? (If “yes,” the item is Acceptable as written. If “no,” proceed to questions 2 and 3.)
2. What is the potentially biased or offensive content and the group(s) affected?
3. Can content be edited to remove potential biased or offensive content? (If so, please offer an edit.)

Bias review committee members viewed printouts of the items as well as note taker select functional versions from IBIS that were projected on a screen, allowing committee members to see representations of how tasks were presented and how students were to select responses. ETS assessment specialists facilitated discussion of each item with reference to the guiding questions.

Ideally, the content committee reviews would have been completed well ahead of the bias committee reviews, allowing sufficient time for content committee revisions to be entered in the item bank prior to the bias committee review. As timelines did not allow for this, the events were held on an overlapping schedule, and ETS created a workflow that supported staggered handoffs. While the content reviews were conducted from August 18 until noon on August 22, the bias review committees began their work on August 20 and completed work by the end of the day on August 22. In order to allow the bias review committees time to consider the revisions that had been recommended during content reviews, the content review committees delivered spreadsheets of their revisions via password protected flash drives. After bias review committee members had the opportunity to review items independently, the note taker read comments from both the TMT and the content committee for the bias review committee to consider. The facilitator moderated discussion regarding TMT observations, committee member observations, or observations from bias review committee members and helped the committee reach consensus on overall decisions whether to accept an item as written, accept an item with revisions, or reject an item. When an item was accepted with revisions, the facilitator helped the committee reach consensus regarding the description of the revision. The note taker recorded the group's decisions and descriptions of any needed revisions.

#### *3.3.4.E Process for Handling Issues Outside the Scope of Reviews*

During the committee meetings, participants sometimes had questions, observations, or concerns regarding issues that were outside the scope of the committee reviews. This section describes the process for handling committee members' comments regarding issues outside the scope of the reviews.

With a large number of items to review within a short span of time, it was important for committee members to remain focused on reviewing and evaluating the quality of the items. Facilitators were responsible for helping the committees retain their focus. However, because the opinions of committee members—educators from consortium states—were valued, issues that were outside the scope of the guiding questions for content or bias reviews were recorded on posters. Each committee room was equipped with posters and markers for recording such “parking lot” issues. Questions that could not be answered by the facilitators were answered by ELPA21 or ETS representatives during the committee meetings, when possible. Other issues and concerns were collected at the end of the meetings, compiled in a document by ETS, and shared with ELPA21 for further consideration.

This process of recording issues on posters was successful because it allowed participants to express and share their concerns in an efficient manner. ELPA21 found it useful to have the issues compiled for their consideration. One example of the concerns that was raised at the Kindergarten and Grade 1 levels was that writing skills were to be assessed via computer-based questions only, without having students do any writing on paper. This was further information to support ELPA21's decision to develop additional paper-based writing items for administration to Kindergarten and Grade 1 students.

#### 3.3.4.F Results of Content and Bias Review Committee Meetings

The content and bias review committee meetings had goals related to item quality and to professional development. This section provides the results of the meetings in relation to those goals.

The item quality goals were to ensure that items for the ELPA21 field test are:

- of the highest possible quality;
- aligned to the ELP Standards; the ELA, math, and science practices; and the PLDs;
- free of bias and sensitivity issues.

The content committees reviewed the quality of the items and their alignment in relation to the ELP Standards, the practices, and the PLDs. The content review committees augmented prior work on the alignment of items to the ELP Standards and the practices by reviewing prior decisions from item writers and assessment specialists. The content review committees either confirmed the prior decisions regarding alignment to the ELP Standards and the practices or revised them as necessary. The content review committees made initial determinations regarding the alignment of items to the PLDs under the guidance of the assessment specialists who served as facilitators during the sessions. The bias review committees reviewed the quality of the items and made any revisions needed to ensure that they were free of bias and sensitivity issues.

A total of 2,685 items were reviewed by the content and bias review committees; 1,420 were accepted as written, 1,120 were accepted with revisions, and 145 were rejected (130 rejected by content review committees and 15 rejected by bias review committees). The overall acceptance rate was 95 percent. Note that these numbers are based on the initial tallies compiled immediately after the content and bias committee reviews. As described in Section 3.3.4.G, some items that were not approved at the committee meeting were later revised, based on committee direction, for review and approval by the TMTs. As a result, the totals do not precisely correspond with the final count of items that the TMTs eventually approved for field testing. Tables 8-12, in Appendix A, provide details on the full pool of approved ELPA21 intended operational items by task type (Table 8), by alignment to ELP Standards (Table 9), by alignment to practices (Table 10), and by alignment to PLDs (Tables 11 and 12).

The content and bias review committees succeeded in reviewing all ELPA21 items according to criteria that were approved by the TMTs. The committees succeeded at integrating input from educators across ELPA21 consortium states to enhance the quality of the ELPA21 field test items and ensure that they are aligned to ELP Standards, practices, and PLDs.

In addition to the goal of reviewing test items, the content and bias review committee meetings also had the following professional development goals:

- to inform educators about the quality processes underlying ELPA21 item development;
- to provide an effective professional development experience for educators.

To satisfy the first professional development goal, content and bias review committee members attended presentations that provided overviews of the item development processes and rigorous review processes for ELPA21 items. Their participation in the committee meetings gave educators experience reviewing the quality of items under the consideration of appropriate criteria as well as experience following security protocols for a high-stakes assessment. Although educators are not allowed to communicate the content of any of the items they reviewed as proscribed by the confidentiality agreements that they signed, they were encouraged to share information about the item development and review processes with their colleagues when they reported back to their schools and district offices.

Participant surveys provide a clear indication that the committee meetings fulfilled the second professional development goal of providing an effective professional development experience for educators. A total of 82 educators participated in the committee meetings (61 content and 21 bias review committee members) and 53 of them completed the online participant survey (40 content and 13 bias review committee members). The first part of the survey provided a series of statements to which participants responded by selecting one of four choices: Strongly Disagree, Disagree, Agree, or Strongly Agree. When asked to reflect on the general training presentation, a total of 51 (96 percent) of the respondents agreed or strongly agreed with the statement, “The presentation helped me understand the mission and goals of ELPA21.” When asked to think about the grade-level committee meetings, total of 50 (94 percent) of the respondents agreed or strongly agreed with the statement, “The materials provided to support the committee meeting discussions were clear and well organized.” A total of 52 (98 percent) agreed or strongly agreed that the committee meeting facilitators were knowledgeable. A total of 51 (96 percent) of the respondents agreed or strongly agreed that they were able to provide meaningful input to the ELPA21 test items. When asked how they would describe the overall event, 44 (83 percent) of the respondents described it as excellent, 9 (17 percent) described it as good, and none described it as fair or poor. All 53 of the respondents indicated that they would participate in such an event again.

Respondents had an opportunity to submit any comments that they wished to share. The positive tone of the comments that were provided corresponded with the positive responses to the SR questions. The following are representative comments provided by committee members:

“This was such a positive experience. I felt valued and appreciated in all aspects from selection to information received from the coordinators, to the actual conference. Everything was top-rate and it was one of the most enjoyable experiences I have had.”

“I came to this event with low expectations about my level of excitement about the work, but I was so wrong! I was very engaged in the work, and I really feel like my input and expertise were valued and important. I didn't want the week to end. The event was extremely well organized and every member of ELPA21, ETS, CCSSO and other leaders were so knowledgeable and available to answer our questions and talk about the work.”

“I feel like I really made a difference - that my suggestions, comments, concerns, and ideas were acknowledged and respected. I learned so much about the test development process. I feel equipped to lead my district through the new ELP Standards and accompanying assessment.”

The respondents indicated that they valued the experience because they learned more about mission and goals of ELPA21; they participated in discussions that were supported by knowledgeable facilitators and clear, well organized materials; and they were able to recommend appropriate revisions to ELPA21 field-test-ready items. The survey responses indicated that the content and bias review committee meetings provided a successful opportunity for participants to develop as educators and enhance the quality of the ELPA21 field test item pool.

#### *3.3.4.G Resolution WebEx Meetings with Task Management Teams*

After the content and bias review committee meetings, the remaining major step in finalizing the content of the ELPA21 field test pool was a series of meetings, held via WebEx, to resolve and agree on revisions with the IAD TMT, the AAA TMT, and invited additional stakeholders from CCSSO and ELPA21.

Two WebEx meetings were held, each lasting approximately three hours. The ETS Core Team began by presenting a summary of the results of the content and bias review committee meetings reporting on such issues as number of participants, processes for facilitating the committee meetings (and ensuring security of the item pool), results of the meeting (in terms of numbers of items accepted as written, accepted with revisions, and rejected), and comments made by participants in their evaluations.

The ETS Team then presented to the IAD and AAA TMTs the recommendations for revisions to items for each grade or grade band, discussing overall themes and categories of recommended revisions; the disposition of comments made by TMT reviewers in their pre-committee meeting reviews; and examples of items accepted as written, accepted with revisions, and rejected. For a relatively small group of items that had been rejected in the committee meetings, ETS also proposed revisions to address the concerns raised by the committees and presented those to the TMTs. For example, some items required brand new art, and others required replacement items for item sets. The TMTs reviewed and approved these revisions or new items, modestly increasing the number of items available for field testing.

After the ETS presentations and discussions among all parties in the meeting, the IAD and AAA TMT members then voted on acceptance of the revision process for the items presented. Revisions for Kindergarten, Grade 1, and Grades 2-3 were presented in the first day’s WebEx; revisions for Grades 4-5, Grades 6-8, and Grades 9-12 were presented in the second day’s WebEx. In all cases, the TMT members voted to accept the revision process and expressed approval of and satisfaction with the review and resolution process.

A total of 2,469 intended operational items and 150 experimental items were accepted. Tables 8–12 in Appendix A provide details on this item pool.

After the WebEx resolution meetings were concluded, the ETS Team turned to tasks associated with readying the main ELPA21 item pool for handoff to the field test delivery vendor. Key steps here included applying all changes to test content recommended by the content and bias review committees (including changes to graphics), professional recording of audio as needed for the listening domain and aspects of other domains, and reviewing and executing quality control measures for the metadata coded to each item.

#### *3.3.4.H Audio Recording*

Developing high-quality audio recordings is essential to an English language proficiency assessment. Audio recordings constitute the stimulus for listening items. For ELPA21, audio recordings were also used to supplement text for certain reading, speaking, and writing items and directions. ETS was able to draw upon established experiences, practices, and networks to recruit appropriate voice actors and develop high-quality recordings.

During the development of items, it became apparent that it would not be practical for educators to listen to all audio recordings as they reviewed items at the content committee meetings. Plans were made to develop a set of demonstration audio recordings that would be reviewed by the TMTs and then by content committee members. Based on discussions with TMT representatives, the following set of guidelines was developed.

#### **Recording of Listening Stimuli**

All listening stimuli for ELPA21 will be professionally recorded in a manner that is appropriate to their intended purposes, gathering evidence about the listening proficiency of students in the target population.

- All stimuli will be recorded at a realistic (though not overly fast) pace, and with natural prosody.
- All voices will represent fluent speakers of standard American English; no discernible geographic accents will be used.
- Professional voice actors will be used for all roles, including the roles of children.
- Adult voice actors will not voice the roles of children; child voice actors will be used. All child voice actors will produce clear, easily comprehensible speech.
- The voice used as narrator will not also be used to deliver dialogs, monologs, or other stimulus content.

*(ELPA21 Editorial Style Guide, 2014, p. 46)*

Sample audio recordings developed in accordance with these guidelines were also played for educators during the content committee review meetings, allowing educators to evaluate the speed of delivery and the general quality of the recordings.

In response to educator feedback that came during the item development training and the content review committee meetings, the directions, stems, and options were recorded for certain domains. Speaking directions and stems were recorded so that students would not have to rely on reading skills to comprehend the prompt. Similarly, writing directions, stimuli, and stems were

recorded for most item types in each grade or grade band to ensure that students would not need to rely on reading skills to comprehend the prompts. For lower grade bands, directions were recorded for reading items. The text of Kindergarten reading passages and items was accompanied by audio recordings since the ELP Standards do not require Kindergarten students to read text independently for comprehension.

One challenge of creating high-quality recordings was to identify appropriate voices for Kindergarten, Grade 1, and Grades 2–3. The stimuli and prompts required voices that sounded like young children of that age, but children of that age are not typically able to read and narrate script so that it sounds like fluent, spontaneous speech. This issue was resolved by using two groups of voice actors: one group of actors between the ages of 9 to 11 read the scripts for Kindergarten to Grade 5, while another group of actors aged 14 to 17 provided the voices for the Grade 6 to 12 test materials. Multiple actors were used to reflect the variety of roles that needed to be filled in each situation. In addition, enough actors were hired to avoid using the same actor for all teacher or student roles within a given task type. Using a variety of voices allows students to hear a new set of voices within each setting, enhancing the authenticity of conversations and preventing students from being distracted by hearing the same voice again and again.

After guidelines for selecting voice actors and creating quality audio recordings had been established, they were used to develop audio recordings for the pool of ELPA21 items. As soon as revisions from the content and bias review committees had been entered in IBIS, audio recordings were developed for listening items as well as specific items and directions from other domains.

Assessment specialists served as directors for the audio recordings, ensuring that high-quality audio recordings with appropriate delivery were developed. The ETS Core Team member directing the recording assisted actors with their delivery to ensure that pronunciation was correct and consistent and that grade-appropriate pacing, tone, and pitch were used. The director also monitored emphasis so that all lines sounded natural and important (tested) concepts were appropriately highlighted. After recording sessions, the audio files were securely transferred from the recording studios to ETS, where they were uploaded to the appropriate items in IBIS. Assessment specialists then proofed the audio files against the scripts to ensure that the recordings were accurate and appropriate. Arrangements for re-recordings were made when needed.

The process of demonstrating sample recordings to TMTs and educators allowed ETS to collect feedback and establish guidelines for developing high-quality, age-appropriate audio recordings for the full body of ELPA21 items.

### *3.3.5 Accessibility for Students with Disabilities*

#### *3.3.5.A Employing Principles of Universal Design*

A key goal of the ELPA21 assessment system is to maximize accessibility for all students, including students with disabilities. Ways of achieving this goal include following principles of

Universal Design in the item design process, following principles of Universal Design in the item writing and review process, evaluating all items for overall accessibility, embedding accessible content using the APIP standard to items that can be rendered accessible via the accommodations specified by APIP, and developing “twins” for items judged not to be accessible to students with visual impairments in their form as developed for field testing.

All ELPA21 items were designed and written following the principles of Universal Design. Universal Design was incorporated into the process in multiple ways. Item writers and internal and external reviewers were trained in Universal Design for assessment including the following as outlined by Thompson, Johnstone, and Thurlow (2002):

1. Inclusive assessment populations
2. Precisely defined constructs
3. Maximally accessible, non-biased items
4. Simple, clear, and intuitive instructions and procedures
5. Amendable to accommodations
6. Maximum readability and comprehensibility
7. Maximum legibility

Beginning with the early stages of task design, consideration was given to accessibility. Each task type was reviewed for overall accessibility and suggestions were made for ways to alter task types to be more accessible for ELLs with disabilities. A number of computer-based tasks were discussed in the early design phase for ELPA21, but only task types that were potentially amenable to the APIP process were selected for use.

Artists received art guidelines with specifications on the creation of maximally accessible illustrations and other graphics. (See Section 3.2.6 Editorial Style Guide and Graphics Style Guide for more details on these guidelines.)

### *3.3.5.B Collaboration with Accessibility, Accommodations, and Administration Task Management Team*

The ELPA21 governance structure includes an AAA TMT. The AAA TMT consists of experts with extensive educational research and assessment industry experience and includes state education agency members nominated by the ELPA21 consortium states. The AAA TMT was regularly consulted throughout the item design and development process, with representatives of the ETS Team participating in the AAA TMT’s regular meetings. The AAA TMT provided official input on all major deliverables in the ECD process. This input included reviews of Item Specifications, contributions to item writer training materials, participation in item writing trainings, pre-committee reviews of items, and participation in the content and bias review meetings. Having the ongoing input of these experts was important in keeping accessibility considerations at the forefront of item development. In addition, the AAA TMT was part a key part of the group that approved the twinning solution proposed by ETS and further described in section 3.3.5.D.

### *3.3.5.C Evaluation of Item Accessibility and Use of Accessible Portable Item Protocol*

As item writers and reviewers completed their work on the items, items were sent to a workflow step where ETS's Alternate Test Format group reviewed each item and evaluated accessibility. For ELPA21, in addition to a general accessibility review for students with disabilities, a separate accessibility review focused solely on the accessibility of items for students with visual impairments (including blindness/low vision). Each item in the pool was evaluated to determine whether the item was accessible, accessible with adaptation (and, if so, the type of adaptation), or inaccessible. It should be noted that in order to reduce the linguistic load of the assessment for young learners and ELLs who are non-readers or who are at a lower level of language proficiency, visual images were an important part of the content of the assessment. For items with visual images, a determination was made as to whether text descriptions (alternative text or "alt-text") could be written so the text could then be voiced by a text-to-speech engine or presented as a figure description for Braille delivery according to student needs. In cases where a text description of a graphic would not interfere with what was being measured, alt-text was written. These accessibility determinations and alt-text were then reviewed by members of the ETS Core Team to confirm that the alt-text descriptions were written in grade-appropriate language.

Once the item was reviewed in full, accessible content was then generated for all items deemed to be accessible. The accessible content was structured as defined by the APIP v1.0 standard. The ELPA21 item pool as delivered for field testing contains 1,178 items with APIP-structured accessible content.

### *3.3.5.D Braille-Ready Items for Paper Delivery*

Although accessible items for students with visual impairments can be delivered via computer through a refreshable Braille keyboard, some states may still need to administer paper Braille forms. For that reason, ETS was asked to ensure a suitable pool to support field testing a paper version of a Braille form. This included the need to develop "twins" for non-accessible items. Twinning is an accessibility solution to create accessible items for students whose disabilities may impact their opportunity to access the content. Twinning can be described as a process where original items are identified as non-accessible for the target population, and the non-accessible item is then "revised" with a series of action steps specific to the item type. The result is the twinned item type that provides an opportunity to measure the students' skills. For example, a listening item with picture options was found to be inaccessible to students with visual impairments. The twin developed for this task type featured manipulatives, i.e., tangible, three-dimensional objects, instead of the visual picture options.

Because the youngest learners in the ELPA21 test-taking population (especially in Kindergarten and Grade 1) and students with late-onset visual impairments including blindness cannot be assumed to be Braille readers yet, significant care was taken to ensure that the new twin types did not introduce any construct irrelevant variance into any of the four domains. That is, for the listening domain, answering questions accurately should depend only on a student's level of

listening proficiency and not on the ability to read Braille. Therefore, in addition to typical twin solutions such as reformatting, additional solutions were found including the use of manipulatives, creation of tactile graphics, physical response, new presentation, and new tasks. Also, twin items in the writing domain allow for responses that could include a range of response possibilities, e.g., specialized paper suitable for students with low-vision, Braille word tiles, slate and stylus, Braillewriter. This response type would depend on the test-taker's Individualized Education Program (IEP) or 504 Plan as well as a formal policy decision from the ELPA21 Consortium and/or individual state policy decisions.

The ELPA21 twin items were developed and then reviewed internally through ETS's standard review procedures and then by the IAD and AAA TMTs. In addition to creating twin items, any additional items in sets containing twins were entered into the item bank and coded for paper delivery. This resulted in pool of 415 twins and associated items suitable for administration in a paper Braille format for test-takers with visual impairments. Table 14 in Appendix A provides details on the distribution of these twins and associated items.

### *3.3.6 Finalization and Export of Item Pool*

After all item reviews—ETS internal reviews, TMT reviews, and reviews by the content and bias review committees—had been completed and resolved, the focus of the item development work turned to ensuring that all of the decisions made in those reviews had been implemented accurately and consistently, and to preparing the item pool for export to the field test delivery vendor.

Based on the resolution of comments in the various reviews, any item in the pool might need to have changes of one or more of the following types applied:

- Changes to art or other graphics;
- Changes to language in the item (which might affect the item text as presented and/or the audio script);
- Changes to metadata.

The application of these change was executed in a logical sequence; because changes to art and graphics are more time-consuming, they were given first priority. All changes were subject to rounds of independent checking to ensure that they had been executed accurately and that any inconsistencies were noted and resolved. At the same time that art, content, and metadata changes were being made, audio files were recorded, proofed for accuracy, and embedded in designated locations within the structured item content as described in Section 3.3.4.H.

When all updates had been made to the item, a technical review was conducted to ensure that the items were ready for export. In the technical review, ETS systems and capabilities staff inspected the XML underlying the item entry; they also conducted

several automated validations to ensure that item content was presented in a manner consistent with the Editorial Style Guide and Graphics Style Guide. For items with accessible content, the APIP tags were then generated based on the previously authored accessible content, and the items were made ready for exporting. The items were exported from IBIS, an extensively tested utility that has received all available APIP conformance certifications from the IMS Global Learning Consortium.

There were 24 export batches of field-test-ready items, consisting of the six grades or grade bands and the four domains (listening, reading, speaking, and writing). Quality control reviews were executed to make sure that all of the appropriate items were in a given batch. The metadata were reviewed to make sure that appropriate fields had required data, the data matched the input from committees, and the data were internally consistent. After the metadata reviews, ETS staff made sure that a snapshot of each item was generated and included in the export. The batch was then exported out of IBIS. At the time of export, the IBIS system makes sure that the XML is valid and conforms to Question and Test Interoperability (QTI) standards. A final set of materials for delivery, consisting of the exported and zipped QTI packages and metadata export Excel files for each export batch, were then posted on the field test delivery vendor's Secure File Transfer Protocol (SFTP) site.

The finalization and export of an item pool are always somewhat complex and painstaking processes, because it is at this point that the iterative nature of the ongoing development processes comes to an end, and all item content must be in complete and accurate form, ready for field testing. In the case of ELPA21, the complexity of these processes was increased somewhat by the compressed overall timeline, which meant that some conceptual decisions, including finalization of metadata fields and values, continued to be made relatively late in the item development process. In addition, the project schedule required that the evaluation of the item accessibility and APIP (described in Section 3.3.5.C) and the preparation of Braille-ready items for paper deliver (Section 3.3.5.D) occur simultaneously with preparations of the main item pool for export.

Given the importance of portability to ELPA21, some limits to the APIP standard are also worth mentioning. While APIP specifies fairly detailed information about item content, it does not document decisions related to art or audio files (e.g., the size or type of files, specifications about how art should appear), how the item should be displayed (e.g., how the passages is displayed with the item), or aspects of text formatting (e.g., use of bold font in item stems). To maximize the likelihood of a transfer from vendor to vendor be executed smoothly, it is important for details of this sort to be worked out early in the process so that the exchanges can be made smoothly and rework will not be needed.

After the item export, the item pool was reviewed by the field test delivery vendor and several discussions were held to resolve questions and potential issues related to the items

as they were readied for field test delivery. In some cases, it was decided that it was appropriate for ETS to revise item XML to use format tags that could be accommodated by the field test delivery vendor's system. For example, several ELPA21 task types depend on a "drag and drop" functionality in which students move a "source" image or piece of text to a "target." In some cases, the ETS system handles the source as text while the field test delivery vendor's system handles it as art, and ETS re-authored the items to meet the requirements of the vendor's system. In other cases, the sizing of the sources and targets was not appropriately consistent, and ETS re-authored the items to provide more consistent sizing.

### *3.3.7 Additional Scope of Work*

During the course of the ELPA21 item pool design and development effort described above, ETS was also contracted to take on two additional pieces of work. Section 3.3.7.A and 3.3.7.B describe this work.

#### *3.3.7.A Cognitive Laboratory Study*

During the course of the item development effort, ELPA21 requested that ETS conduct a small-scale cognitive laboratory study to examine the quality of the items with a focus on new TE features. Specifically, the purpose of the study was to examine student interaction with new ELPA21 item types and use of technology features during the computer-based testing so that issues could be addressed prior to the 2015 field test, if possible, and before operational testing in 2016. The areas of investigation focused on the clarity of item directions, the usability of technology features (e.g., drag and drop, hotspot, recording functions, mouse control, keyboarding), and accessibility features.

A total of 91 students in Grades K–3 and Grades 6–8 were recruited from two elementary schools and one middle school in two urban districts in Seattle, Washington. Due to time and resource constraints, not all of the grade levels were included in the study. The participants included current ELLs at different ELP levels, fluent English-speaking students (non-ELLs), and ELLs with an IEP or a 504 Plan (i.e., ELLs with disabilities). The first or home language of the ELLs in the sample included Spanish, Russian, and Vietnamese.

Overall, the study yielded promising results for the quality of the item types and use of technology features. The study yielded the following findings:

- Most students were able to quickly learn to use technology features and complete the TE item types on a computer when clear directions were provided.
- Students were highly engaged and interested in the tasks. This observation was particularly evident in younger grades.
- Students at Grades K and 1 needed individual assistance in understanding directions and/or in using the mouse or other technology to complete the TE tasks during the computer-based testing.
- Students with a beginning level of ELP needed extra support in understanding the directions across all the grade levels.

Based on these findings, recommendations and suggestions were provided for future development or administration of the operational ELPA21 assessments, including the following:

- A tutorial and sample practice items should be provided to the students prior to the testing.
- Clear and explicit directions must accompany each item type.
- The construct measured in each item type should be clearly defined for computer-based testing.
- Accessibility tools designed to support all students' access to the test content should be designed to work with innovative TE item types and the computer platform; all tools must be demonstrated to all students.
- Individual assistance should be provided during the test administration for students in Kindergarten and Grade 1.
- Item formatting should be clear and intuitive and should not require unnecessary use of technology navigation features.
- Technical requirements should include detailed specifications about the equipment and must be tested prior to field and operational testing.

Results of the cognitive laboratory study demonstrated the need for further empirical research to monitor the appropriate use of technology features and provide useful information to enhance the validity and technical quality of the ELPA21 assessments.

Due to the timing of the study, which took place after field-test-ready items had already been finalized, the opportunity to implement changes to the intended operational items in advance of the field test based on the findings of the study was limited. However, steps were taken by the IAD TMT and field test delivery vendor to address issues based on student interactions with item types by adding emphasis to appropriate areas in the administration manuals, addressing problematic areas in the interactive online sample items used to familiarize students with the assessments, and making adjustments to the test delivery platform.

A full report on the study is available under the title, "Investigating the Usability of Technology-Enhanced Assessment Items during the ELPA21 Assessment Development Process" by Mikyung Kim-Wolf, Danielle Guzman-Orth, and Jennifer Wain, all of ETS.

### *3.3.7.B Paper-Based Writing Tasks for Kindergarten and Grade 1*

One of the requirements of the ELPA21 contract was that all items be designed to be delivered in a computer-based format. While computer-based delivery has significant advantages, there were discussions from early in the task design process that computer-based tasks might not be sufficient to assess the standards related to writing skills for students in the lowest grades who generally do not yet have keyboarding skills. ETS designed several computer-based tasks of writing skills for these students, and these task types were approved by the TMTs; however, there was concern that ELPA21 might be better served by also including direct measures of writing skills for these students, which would necessitate the use of a format allowing students to respond by putting pencil to paper. This interest in paper-based writing tasks was strengthened

by the input of the Kindergarten content committee, who provided a clear opinion during content review that computer-based tasks alone were not sufficient to assess the writing standards, and paper-based tasks should also be included.

ELPA21 issued a change order to ETS to design and develop a supplementary set of paper-based tasks aligned to the writing sub-claims for Kindergarten and Grade 1. ETS started work on this task immediately after the delivery of the main item pool. This design and development effort followed the same ECD approach as for the main item pool, with the requirement that the items be amenable for computer-based delivery and response being replaced with the requirement that these new items be amenable to paper delivery and response.

The IAD TMT reviewed an initial draft of the Item Specifications and sample items, and ETS revised those documents based on TMT input before developing the items themselves. ETS delivered a total of 47 items spread across five Kindergarten task types and four Grade 1 task types. In agreement with ELPA21, ETS delivered the draft paper-based writing items in manuscript form; that is, ETS delivered the text of the items along with graphics files, metadata, Item Specifications, draft scoring rubrics, and directions for administration, but did not enter the items into any item banking system.

## References

- Council of Chief State School Officers. (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: CCSSO.
- Council of Chief State School Officers. (2014). *English language proficiency standards with correspondences to K–12 English language arts (ELA), mathematics, and science practices, K–12 ELA standards, and 6-12 literacy standards*. Washington, DC: CCSSO.
- Mislevy, R. J. (2011). *Evidence-Centered Design for simulation-based assessment. CRESST Report 800*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Mislevy, R.J., Steinberg, L.S., and Almond, R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.
- Thompson, S.J., Johnstone, C.J. & Thurlow, M.L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

APPENDIX A: Additional Tables of Items Produced

**Table 8: Intended Operational Items by Task Type**

Domain	Task Type	Grade or Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
<b>Listening</b>	Academic Debate					17	15	32
	Academic Lecture and Discussion						18	18
	Academic Lecture or Discussion					11		11
	Follow Instructions	54	20	16	20	6		116
	Interactive Student Presentation				14	13	12	39
	Listen and Match	58	50	43	39	38	39	267
	Listen for Information				12	11	16	39
	Long Conversation	18	12	12				42
	Read-Aloud Story	18	16	18				52
	Short Conversation	3	10	8	12	22	17	72
	Student Discussion				13			13
	Teacher Presentation	18	15	20				53
	Teacher Presentation: Read Aloud				15			15
<b>Listening Total</b>		<b>169</b>	<b>123</b>	<b>117</b>	<b>125</b>	<b>118</b>	<b>117</b>	<b>769</b>
<b>Reading</b>	Argument and Support Essay Set					18	21	39
	Discrete Items					26	39	65
	Extended Informational Set				24	19	19	62
	Extended Literary Set				26		30	56
	Extended Literature Set					16		16
	Informational Set	15	23	24				62
	Literary Set		22	22				44
	Match Picture to Word and Sentence				32			32
	Procedural Text		19	15				34
	Read and Match	58	44	19				121
	Read for Details		6	5				11
	Read-Along Sentence		22	20				42
	Read-Along Story	21						21
	Short Correspondence	21	20	30				71
	Short Correspondence Set				16			16
	Short Informational Set				16	16	22	54
	Short Literary Set				11		24	35
	Short Literature Set					17		17
Word Wall	30						30	
<b>Reading Total</b>		<b>145</b>	<b>156</b>	<b>135</b>	<b>125</b>	<b>112</b>	<b>155</b>	<b>828</b>
<b>Speaking</b>	Analyze a Visual				14			14
	Analyze a Visual and a Claim					10		10
	Analyze a Visual and a Claim Argument						10	10

Domain	Task Type	Grade or Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
	Classroom Tableau	36	35	25				96
	Compare Pictures			11	6	5	6	28
	Conversation		15	18	24			57
	Language Arts Presentation				18	15	15	48
	Observe and Report	16	4	3	3	3	3	32
	Opinion		14	8				22
	Oral Vocabulary				20	20	20	60
	Picture Description	35	7					42
	Show and Share Presentation	24						24
	Show and Share Questions	12						12
<b>Speaking Total</b>		<b>123</b>	<b>75</b>	<b>65</b>	<b>85</b>	<b>53</b>	<b>54</b>	<b>455</b>
<b>Writing</b>	Complete the Story	14						14
	Construct a claim					3	5	8
	Discrete editing tasks				14	9	16	39
	Opinion			6				6
	Picture Caption			25				25
	Respond to a Peer E-mail					7	5	12
	Sentence Builder	20	75	31	19			145
	Storyboard			9	6	5	6	26
	Word Builder	57		22	11			90
	Write an Opinion				7			7
Writing Questions Task				15	15	15	45	
<b>Writing Total</b>		<b>91</b>	<b>75</b>	<b>93</b>	<b>72</b>	<b>39</b>	<b>47</b>	<b>417</b>
<b>Grand Total</b>		<b>528</b>	<b>429</b>	<b>410</b>	<b>407</b>	<b>322</b>	<b>373</b>	<b>2469</b>

**Table 9: Field-Test-Ready Items Aligned to Each English Language Proficiency Standard**

Note: Many items are aligned to multiple standards.

Grade or Grade Band	Domain	Standard									
		1	2	3	4	5	6	7	8	9	10
<b>K</b>	Listening	160	21	8*	8*	18	3		181		8*
	Reading	127	21			51			145		
	Speaking		31	87	19	16				16	118
	Writing			91							91
<b>1</b>	Listening	163	37	12*	8*		4		143		12*
	Reading	156	20						112		
	Speaking		29	75	14	4				4	75
	Writing			75						25	25
<b>2-3</b>	Listening	108	20	4*		19	12		117		
	Reading	135	30	4*					150		
	Speaking		18	39	8					3	36
	Writing		6	87	6					9	78
<b>4-5</b>	Listening	131	66	2*	2*	34	14		61	2*	2*
	Reading	135	61	2*	1*	19	17		54	2*	2*
	Speaking		31	65	13			3		79	85
	Writing		22	43	7			28		58	72
<b>6-8</b>	Listening	114	24	2*	2*	15	24	2*	45	2*	2*
	Reading	104	99	2*	2*	15	16	2*	44	2*	2*
	Speaking		53	53	18	14	11	3		53	53
	Writing		30	15	10			23		23	39
<b>9-12</b>	Listening	102	52	2*		16	26	2*	125		
	Reading	168	61	2*		52	26		168	2*	
	Speaking		15	54	14	10	5	3		3	3
	Writing		20	6	10			26		47	47

\*Denotes experimental integrated items that address both receptive (reading or listening) and productive (writing or speaking) domains.

**Table 10: Field-Test-Ready Items Aligned to Each Practice**

Note: Many items are aligned to multiple practices.

Grade or Grade Band	Domain	English Language Arts						Mathematics							Science							
		EP1	EP2	EP3	EP4	EP5	EP6	MP1	MP2	MP3	MP4	MP5	MP6	MP7	SP1	SP2	SP4	SP5	SP6	SP7	SP8	
<b>K</b>	Writing		91				3															
	Listening	189		6	18		8	1		6			3		1					6	26	
	Reading	145			30		21														66	
	Speaking	107			16		87						37						87		90	
<b>1</b>	Listening	133		8			22	24		8			22							8	49	
	Reading	153				10	22		6				29		9	22	1	10	10		86	
	Speaking	7			4	41	87			1			6						18		53	
	Writing		75				70	25					25								25	
<b>2-3</b>	Listening	50	4			20	59														20	
	Reading	51	4		17	30	82														12	
	Speaking			8		18	36												3		3	
	Writing		39	6			54															
<b>4-5</b>	Listening	114		30	40		12	3		18			3	22		22	12		3		39	
	Reading	104	2	20	99	45	18	45	7	5		46		49		49	46	46	30		101	
	Speaking			14	41	36	85			20			7	14		14				20	85	
	Writing		72	7		28	58			7			14	43	15					7	43	
<b>6-8</b>	Listening	89	2	1	10	1	2	1		1			1			10	5		1	1	14	
	Reading	122	2	3	1		2	2	4								8	1			26	
	Speaking	50		9	4		47		1	9	1		10				11				14	
	Writing	33	20	11		12	12															
<b>9-12</b>	Listening	128		10	10	6	1			12	1		42				23		2	13	127	
	Reading	170		5	131	1	10	16		6			54		10		64			1	170	
	Speaking	30		18	10	11	7			7			28	1			2		17	7	54	
	Writing		36	5		5	36	22		5			39		15		11		5	10	36	

**Table 11: Field-Test-Ready Items Aligned to Each Proficiency Level Descriptor (Part 1)**

Grade or Grade Band	Domain	1.1 1.2 1.3 1.4 1.5					2.1 2.2 2.3 2.4 2.5					3.1 3.2 3.3 3.4 3.5					4.1 4.2 4.3 4.4 4.5					5.1 5.2 5.3 5.4 5.5					
<b>K</b>	Listening	71	93	103	48	18	3	3	18	18		8	8	8	8	8	8	8	8	8	8					18	18
	Reading	70	70	51	42	57			21	21	21											30	51	51	51	51	
	Speaking							19	19	31	31	87	87	87	87	87	19	19	19	19	19	16	16	16	16	16	
	Writing											91	91	91	91	91											
<b>1</b>	Listening	58	119	107	20	13		26	33	5	1	12	12	12	12	12	8	8	8	8	8						
	Reading	44	66	112	112	64			20	20	20																
	Speaking						29	29	29	29	29	75	75	75	75	75	14	14	14	14	14	4	4	4	4	4	
	Writing											75	75	25													
<b>2-3</b>	Listening	22	44	49	8			20	3					4	4							4	8	19			
	Reading	26	19	47	61	14			30					4	4												
	Speaking								18	18		10	13	26	14	3			8	1							
	Writing								6	6		28	53	34	34				6	6							
<b>4-5</b>	Listening	59	17	35	21	4	23	29	5		15	3	3	3	3		3	2	3				6	28	1	4	
	Reading	53	59	22	33		13	50	4	1	6	2	2	2	2		1	1	1	1		6	6	10	3		
	Speaking						31	25	25	25	25	65	59	59	59	59	13	13	13	13	13						
	Writing						22	22	22	7	7	22	34	13	13	13	7	7	7	7	7						
<b>6-8</b>	Listening	46	41	24	17		2	24	2	2	2	2	2	2	2	2	2	2	2	2	2		14	1	1		
	Reading	51	62	49	20	1	56	67	38	19	3	2	2	2	2	2	2	2	2	2	2		6	9	3		
	Speaking						53	53	53	53	53	53	2	2	1	1	18	17	17	17	15	14	14	13	13	3	
	Writing						15	30	30	15	15	15	15	15	15	15	10	10	10	10	10						
<b>9-12</b>	Listening	29	45	21	7			12	12	5	23			2								9	7				
	Reading	16	38	144	120	55			58	59	53			2									11	52	41	19	
	Speaking							9	15	15	8	30	40	35	28	12	3	6	14	14	7	5	9	10	2		
	Writing						20	20	20	5	5	6	6	6	6	6	5	5	10	10	5						

**Table 12: Field-Test-Ready Items Aligned to Each Proficiency Level Descriptor (Part 2)**

Grade or Grade Band	Domain	6.1	6.2	6.3	6.4	6.5	7.1	7.2	7.3	7.4	7.5	8.1	8.2	8.3	8.4	8.5	9.1	9.2	9.3	9.4	9.5	10.1	10.2	10.3	10.4	10.5
<b>K</b>	Listening				3	3						74	96	121	66	18						8	8	8	8	8
	Reading											88	70	51	72	87										
<b>1</b>	Speaking																		16	16	16	106	106	118	118	118
	Writing																					91	91	91	91	91
<b>1</b>	Listening	4	4	4	4	4						41	102	80	11	5						12	12	12	12	12
	Reading																									
<b>2-3</b>	Speaking																		4	4	4	75	75	75	75	75
	Writing																		25			25	25	25	25	25
<b>2-3</b>	Listening		12	1								28	46	45												
	Reading											26	44	50	76	26										
<b>4-5</b>	Speaking																		3	3	3	3	3	3	3	3
	Writing																									
<b>4-5</b>	Listening	9	3	8	4	3						58	1	1	1							3	3	3	3	3
	Reading	5	1	11	1							35		2	19							2	2	2	2	2
<b>6-8</b>	Speaking						3	3	3	3	3											79	79	79	79	79
	Writing						28	28	28	28	28											37	49	28	13	13
<b>6-8</b>	Listening	2	13	9	2		2	2	2	2	2	45										2	2	2	2	2
	Reading	5	3	8	1		2	2	2	2	2	26	29	17	9							2	2	2	2	2
<b>9-12</b>	Speaking																									
	Writing																									
<b>9-12</b>	Listening	1	13	16	7				2			11	40	46	27	1										
	Reading		5	18	16	8							30	83	153	83										
<b>9-12</b>	Speaking			5	5	2		3	3	3	3													3	3	3
	Writing						15	21	26	11	11											34	34	41	16	12

**Table 13: Sample Items by Task Type**

Domain	Task Type	Grade or Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
<b>Listening</b>	Academic Debate					2	3	5
	Academic Lecture or Discussion					4		4
	Follow Instructions	4	2					6
	Interactive Student Presentation				3	3		6
	Listen and Match		3	1				4
	Listen for Information					1	1	2
	Long Conversation	3	3					6
	Read-Aloud Story	3	2	2				7
	Short Conversation		2			3		5
	Student Discussion				3			3
	Teacher Presentation	8	3					11
<b>Listening Total</b>		18	15	3	6	13	4	59
<b>Reading</b>	Argument and Support Essay Set						4	4
	Discrete editing tasks					6		6
	Extended Informational Set				4	6		10
	Extended Literary Set				4			4
	Extended Literature Set					7		7
	Informational Set	3	3					6
	Literary Set		3	4				7
	Procedural Text		2	4				6
	Read and Match		2					2
	Read for Details		2					2
	Read-Along Sentence		1					1
	Read-Aloud Story	3						3
	Short Correspondence	3	2					5
	Short Correspondence Set				3			3
	Short Informational Set					4		4
	Short Literary Set				3			3
	Short Literature Set					4		4
Word Wall	5						5	
<b>Reading Total</b>		14	15	8	14	27	4	82
<b>Speaking</b>	Analyze a Visual				2			2
	Analyze a Visual and a Claim					2		2
	Classroom Tableau	6	5					11
	Compare Pictures			1	1	1		3
	Conversation		3					3
	Language Arts Presentation					3		3
	Observe and Report		1					1
	Opinion		2	1				3
	Picture Description	5	1					6
	Show and Share Presentation	3						3
	Show and Share Questions	2						2
	<b>Speaking Total</b>		16	12	2	3	6	
<b>Writing</b>	Complete the Story	2						2
	Construct a Claim					1		1

Domain	Task Type	Grade or Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
	Discrete Editing Tasks					1		1
	Opinion			1				1
	Picture Caption			1				1
	Respond to a Peer E-mail					1		1
	Sentence Builder		2					2
	Storyboard					1		1
	Word Builder		2					2
	Writing questions task				1			1
<b>Writing Total</b>		2	4	2	1	4		13
<b>Grand Total</b>		<b>50</b>	<b>46</b>	<b>15</b>	<b>24</b>	<b>50</b>	<b>8</b>	<b>193</b>

**Table 14: Item Twins and Associated Items<sup>4</sup>**

Domain	Task Type	Grade or Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
<b>Listening</b>	Academic Debate					3	3	6
	Academic Lecture and Discussion						5	5
	Academic Lecture or Discussion					3		3
	Follow Instructions	4	5	5	6	4		24
	Listen and Match	6	5	5	4	4	3	27
	Listen for Information				3	2	3	8
	Long Conversation	3	3	3				9
	Read-Aloud Story	3	3	4				10
	Short Conversation	1	2	2	3	4	5	17
	Student Discussion				3			3
	Teacher Presentation	3	2	4				9
	Teacher Presentation: Read Aloud				4			4
<b>Listening Total</b>		20	20	23	23	20	19	125
<b>Reading</b>	Argument and Support Essay Set					5	6	11
	Discrete Items					6	6	12
	Extended Informational Set				5	5	5	15
	Extended Literary Set				4		6	10
	Extended Literature Set					4		4
	Informational Set	3	4	3				10
	Literary Set		4	3				7
	Match Picture to Word and Sentence				3			3
	Procedural Text		3	3				6
	Read and Match	6	4	4				14
	Read-Along Sentence			3				3
	Read-Along Story	3						3
	Short Correspondence	3	4	5				12
	Short Correspondence Set				3			3
	Short Informational Set				4	4	4	12
	Short Literary Set				4		4	8
Short Literature Set					4		4	
<b>Reading Total</b>		15	19	21	23	28	31	137
<b>Speaking</b>	Analyze a Visual				2			2

<sup>4</sup> As described in Section 3.3.5, in addition to the twin items themselves, any additional items in sets containing twins were re-entered into the item bank to ensure that a version coded for paper delivery was available. Such “associated items” are included in this table along with the twins.

ELPA21 Item Development Process Report

Domain	Task Type	Grade or Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
	Analyze a Visual and a Claim					2		2
	Analyze a Visual and a Claim Argument						2	2
	Classroom Tableau	6	5	5				16
	Compare Pictures			2	1	1	1	5
	Conversation		3	3	4			10
	Language Arts Presentation				3	3	3	9
	Observe and Report	4	4	4	1	1	1	15
	Opinion		2	1				3
	Picture Description	5						5
	Show and Share Presentation	4						4
	Show and Share Questions	2						2
<b>Speaking Total</b>		<b>21</b>	<b>14</b>	<b>15</b>	<b>11</b>	<b>7</b>	<b>7</b>	<b>75</b>
<b>Writing</b>	Complete the Story	2						2
	Construct a Claim					1	1	2
	Discrete Editing Tasks				2	3	3	8
	Opinion			1				1
	Picture Caption			8				8
	Respond to a Peer E-mail					1	1	2
	Sentence Builder		8					8
	Storyboard			3	2	1	1	7
	Word Builder	9	8	8	5			30
	Write an Opinion				1			1
	Writing Questions Task				3	3	3	9
<b>Writing Total</b>		<b>11</b>	<b>16</b>	<b>20</b>	<b>13</b>	<b>9</b>	<b>9</b>	<b>78</b>
<b>Grand Total</b>		<b>67</b>	<b>69</b>	<b>79</b>	<b>70</b>	<b>64</b>	<b>66</b>	<b>415</b>

## APPENDIX B: Educator Participants, Demographic Information, and Expertise

### **Educators Who Participated in Passage Reviews**

- Karen Anderson, ELP/ELD Teacher/Specialist, Auburn-Washburn Unified School District (USD) 437, Kansas
- Maria Bartemes, ELP/ELD Teacher/Specialist, Johnston Community School District, Iowa
- Irina Blekhman, ELP/ELD Teacher on Special Assignment, Reynolds School District, Oregon
- Becky Boyett, ELP/ELD Teacher/Specialist, Rogers Public Schools, Arkansas
- Marty Christie, District ESOL Coordinator, Salina Public Schools USD 305, Kansas
- Melissa Davis, ELP/ELD Teacher/Specialist, Gresham-Barlow School District, Oregon
- Emily Grimes, ELP/ELD Teacher/Specialist, Lincoln Public Schools, Nebraska
- Gulnora Isaeva, Secondary ELP/ELD Teacher Trainer, Omaha Public Schools, Nebraska
- Kasey Johnson, ELP/ELD Teacher/Specialist, Olathe Public Schools USD 233, Kansas
- Cassandra Kennedy, ELP/ELD Teacher/Specialist, Doddridge County School District, West Virginia
- Bethany Martinez, Instructional Facilitator, Wahluke School District, Washington
- Ramona Parrish, ELP/ELD Teacher/Specialist, Ottumwa Community Schools, Iowa
- Ralph Sattazah, District ELL Assessment Coordinator, Fort Smith Public Schools, Alma, Arkansas
- Jason Schimke, ELP/ELD Teacher/Specialist, Beaverton School District, Oregon
- Cheryl Sparks, ESL Teacher Trainer, Omaha Public Schools, Nebraska
- Rebecca Stearns, Regional ELL Resource Specialist, InterMountain Education Service District, Oregon
- Carol Striskovic, ELP/ELD Teacher/Specialist, Mentor Public Schools, Ohio
- Melinda Sturgill, ELP/ELD Teacher/Specialist, Grand Island Public Schools, Nebraska
- Elise Tickner, ELP/ELD Teacher/Specialist, Hood River County School District, Oregon
- Ana Maria Vergara, ELP/ELD Teacher/Specialist, Springfield Public Schools, Oregon
- Sylvia Yoder, ELP/ELD Teacher/Specialist, Newton Public Schools USD 373, Kansas

**Table 15: Demographic Information and Expertise of Educators Who Participated in Passage Reviews**

Category	Responses	Number
State	Arkansas	2
	Florida	0
	Iowa	2
	Kansas	4
	Louisiana	0
	Nebraska	4
	Ohio	1
	Oregon	6
	South Carolina	0
	Washington	1
	West Virginia	1
Gender	Female	19
	Male	2
Ethnicity	Asian/Pacific Islander	2
	Hispanic	1
	Other Hispanic/American Indian	1
	White	16
	Multi-Racial	0
	Prefer Not to Respond	1
Type of School District	Rural	8
	Urban	13
Current Position	Classroom/Content Teacher or General Education Teacher	0
	ELD/ELP Teacher Specialist	15
	District Administrator	1
	District ELL Coordinator	1
	District ESL Coordinator	1
	ESL Teacher Trainer	0
	ELL Instructional Coach	0
	ESL Assessment Coordinator	0
	ELL Assessment Coordinator	0
	Higher Education	0
Other	3	
Years of ELP/ELD teaching and leadership experience	Less than 1 year	0
	1 to 5 years	1
	6 to 10 years	3
	10 to 15 years	9
	16 to 20 years	4
	21 to 25 years	0
	25 years or more	1
Unreported	3	
Highest Level of Education	Bachelor's degree	4
	Master's degree	15
	Doctoral degree	1
	Unreported	1

### **Educators Who Participated in ELPA21 Item Development Training**

- Karen Adams, ELP/ELD Teacher/Specialist, Jefferson County School District 509J, Oregon
- Susan Awad, ELP/ELD Teacher/Specialist, Shawnee Mission School District, Kansas
- Jessica Burchett, ELP/ELD Teacher/Specialist, Marion City Schools, Ohio
- Charlotte Burnham, Bilingual Teacher, Arlington School District, Washington
- Barbara Caples, ELP/ELD Teacher/Specialist, Bellevue Public Schools, Nebraska
- Betsy Cardenas, ELP/ELD Teacher/Specialist, Waldron School District, Arkansas
- Jan Carder, ELP/ELD Teacher/Specialist, Turner Unified School District 202, Kansas
- Brielle Cerven, ELP/ELD Teacher/Specialist, Atlantic Community School District, Iowa
- Kathryn Chase, Student Achievement Specialist, David Douglas School District, Washington
- Laurene Christensen, Research Associate, National Center on Educational Outcomes, Minnesota
- Tracy Conaghan, ELP/ELD Teacher/Specialist, Eugene School District 4J, Oregon
- Mary Cortinas, Teacher, Walla Walla School District, Washington
- Kate Damgaard, District Administrator, Papillion-La Vista School District, Nebraska
- Tammi Davis, ESOL Director, Huntsville School District, Arkansas
- Kara Dodds, ELP/ELD Teacher/Specialist, Dubuque Community Schools, Iowa
- Laurie Essafi, Teacher on Special Assignment, North Clackamas School District, Oregon
- Catherine Glinsman, ELP/ELD Teacher/Specialist, Washington Community Schools, Iowa
- John Gorman, ELP/ELD Teacher/Specialist, North Clackamas School District, Oregon
- Virginia Guynn, ELP/ELD Teacher/Specialist, Berkeley School District, West Virginia
- Lisa Gwin, ELP/ELD Teacher/Specialist, Southeast Local Wayne County District, Ohio
- Kimberly Harrington, ELP/ELD Teacher/Specialist, Hillsboro School District, Oregon
- Leslie Hartley, Teacher, Sumner School District, Washington
- Nicole Hilton, ELP/ELD Teacher/Specialist, Tigard-Tualatin School District, Oregon
- Judy Hobson, Adjunct Instructor, John Brown University and Arkansas Technical University, Arkansas
- Monica Hulubei Piergallini, ELP/ELD Teacher/Specialist, Yakima School District, Washington
- Kendra Latter, Teacher on Special Assignment, Gresham Barlow School District, Oregon

- Krysta Marlow, ELP/ELD Teacher/Specialist, Hermiston School District, Oregon
- Rachelle Marquez, ELP/ELD Teacher/Specialist, Rogers Public Schools, Arkansas
- Lesley Morgan, ELP/ELD Teacher/Specialist, Harrison County Schools, West Virginia
- Sherry Moural, ELP/ELD Teacher/Specialist, Fremont Public Schools, Nebraska
- Sidni Musick, ESL Adjunct Instructor, Garden City Community College, Kansas
- Sarah Nolan, ELD Program Coordinator, Central Kitsap School District, Washington
- Wendy Orloff, ELP/ELD Teacher/Specialist, Beaverton School District, Oregon
- Ellen Riggs, ELP/ELD Teacher/Specialist, Seattle Public Schools, Washington
- Jamie Romburg, Clover Park School District, Washington
- Laura Salem, K–12 ELL Curriculum Specialist, Lincoln Public Schools, Nebraska
- Myrna Salinas, ELP/ELD Teacher/Specialist, West Linn-Wilsonville School District, Oregon
- Marcia Sanders, ELP/ELD Teacher/Specialist, Seattle Public Schools, Washington
- Mary Savage, ELL Specialist, Gonzaga University, Washington
- Tammy Scarborough, ELP/ELD Teacher/Specialist, Raleigh County School District, West Virginia
- Laetitia Schreier, ELP/ELD Teacher/Specialist, Morrow County School District, Oregon
- Yoshiko Schulz, ELP/ELD Teacher/Specialist, University Place School District, Washington
- Vitaliy Shyyan, Research Associate, National Center on Educational Outcomes, Minnesota
- Amanda Smith, ELP/ELD Teacher/Specialist, Nyssa School District, Oregon
- Carolyn Stearns, District ELL Coordinator, Perry Local School District, Ohio
- Anna Stewart, ELP/ELD Teacher/Specialist, Toledo Public School District, Ohio
- Melanie Stuart-Campbell, Project Coordinator, University of Kansas
- Ann Tessier, ELP/ELD Teacher/Specialist, Davenport Community Schools, Iowa
- Lee Ann Thomas, Kansas City Public School District, Kansas
- Holly Tracy, ELL Instructional Coach, Lincoln Public Schools, Nebraska
- Katie Trimble, ELP/ELD Teacher/Specialist, South Sioux City Community School District, Nebraska
- Richelle Vining-Gonzalez, ELP/ELD Teacher/Specialist, Shelton School District, Washington
- Julie White, ELP/ELD Teacher/Specialist, Wichita Public Schools, Kansas

- Molly Williams, ELL Instructional Coach, Lincoln Public Schools, Nebraska

**Table 16: Demographic Information and Expertise of Educators Who Participated in Item Development Training**

Category	Responses	Number
State	Arkansas	4
	Florida	0
	Iowa	5
	Kansas	6
	Louisiana	0
	Minnesota	2
	Nebraska	6
	Ohio	4
	Oregon	13
	South Carolina	0
	Washington	11
	West Virginia	3
	Gender	Female
Male		2
Ethnicity	Asian/Pacific Islander	1
	Hispanic	2
	Other Hispanic/American Indian	1
	White	43
	Multi-Racial	2
	Prefer Not to Respond	5
Type of School District	Rural	21
	Urban	33
Current Position	Classroom/Content Teacher or General Education Teacher	2
	ELD/ELP Teacher Specialist	38
	District Administrator	3
	District ELL Coordinator	1
	District ESL Coordinator	0
	ESL Teacher Trainer	0
	ELL Instructional Coach	2
	ESL Assessment Coordinator	0
	ELL Assessment Coordinator	0
	Higher Education	4
Other	4	
Years of ELP/ELD teaching and leadership experience	Less than 1 year	0
	1 to 5 years	7
	6 to 10 years	16
	10 to 15 years	14
	16 to 20 years	5
	21 to 25 years	0
	25 years or more	0
	Unreported	12
Highest Level of Education	Bachelor's degree	1
	Master's degree	43
	Doctoral degree	7
	Unreported	3

### **Educators Who Participated in Content Review Committees**

- Krista Anderson, ELP/ELD Teacher/Specialist, Elkhorn Public Schools, Nebraska
- Digna Artiles, ELP/ELD Teacher/Specialist, Longview School District, Washington
- Clara Beas-Fitzgerald, ELP/ELD Teacher/Specialist, Morrow School District, Oregon
- Dale Belcher, Teacher, Independence USD 446, Kansas
- Jenny Billingsley, ELP/ELD Teacher/Specialist, Rogers Public Schools, Arkansas
- Jill Brady, ELP/ELD Teacher/Specialist, Southeast Polk School District, Iowa
- Gwen Brewster, Content Teacher, Sioux City Community School District, Iowa
- Kim Brockman, ELP/ELD Teacher/Specialist, Shawnee Mission School District, Kansas
- Beth Bryce, ELP/ELD Teacher/Specialist, Washington Community School District, Iowa
- Amorina Christensen, ELP/ELD Teacher/Specialist, Spokane Public Schools, Washington
- Laurene Christensen, Research Associate, National Center on Educational Outcomes, Minnesota
- Kristina Collins, ELP/ELD Teacher/Specialist, Klamath County School District, Oregon
- Julia Correia, Director, ESL Graduate Academy, Henderson State University, Arkansas
- Rachel Diemer, Classroom/Content Teacher, Sioux City School District, Iowa
- Mary Donnelly, ELP/ELD Teacher/Specialist, Medford USD 549C, Oregon
- Wendy Dunham, ELP/ELD Teacher/Specialist, Nordonia Hills School District, Ohio
- Pam Erixon, ELP/ELD Teacher/Specialist, Millard Public Schools, Nebraska
- Gretchen Fleming, District ELL Manager, Edmonds School District, Washington
- Patty Fong, Title III Consultant, Hamilton County Educational Service Center, Ohio
- Marilyn Gilberts, ELP/ELD Teacher/Specialist, Battle Ground School District, Washington
- Jennifer Gilliland, ELP/ELD Teacher/Specialist, Centennial School District, Oregon
- Kristen Graff, ELP/ELD Teacher/Specialist, Salina USD 305, Kansas
- Erin Gripper, ELP/ELD Teacher/Specialist, Marion County School District, West Virginia
- Michael Grubic, ELP/ELD Teacher/Specialist, North Clackamas School District, Oregon
- Shelly Hamness, ELP/ELD Teacher/Specialist, David Douglas School District, Oregon
- Shonda Haught, ELP/ELD Teacher/Specialist, Maize USD 266, Kansas
- Leah Hinkle, ELP/ELD Teacher/Specialist, Greater Albany Public School District, Oregon

- Margaret Ho, Program Director, ELPA21 Sustainability Planning
- Anne Hubbell, ELL Instructional Coach, Lincoln Public Schools, Nebraska
- Eric Johnson, Professor of Bilingual/ESL Education, Washington State University Tri-Cities, Washington
- Joan Johnston Nelson, Program Supervisor, Migrant and Bilingual Education, Washington State Office of Superintendent of Public Instruction
- Sheree LeDoux-Leos, ELP/ELD Teacher/Specialist, Portland Public Schools, Oregon
- Sheila Litke, ELL Coordinator, Garden County Schools, Nebraska
- Alan Lytle, Public School Program Advisor, Arkansas Department of Education
- Juvy Mangulabnan, ELP/ELD Teacher/Specialist, Garden City Public Schools USD 457, Kansas
- Suzanne McPherson, District Administrator, Fort Smith Public Schools, Arkansas
- Itandewi Mendoza, ELP/ELD Teacher/Specialist, Harvard Public Schools, Nebraska
- Alicia Miguel, Director of ESL, Kansas City Public Schools, Kansas
- Julie Myers, School Principal, Lexington Public Schools, Nebraska
- Kim Myers, Educational Consultant, Southwest Plains Regional Service Center, Kansas
- Judy Pehrson, ELP/ELD Teacher/Specialist, South Sioux City School District, Nebraska
- Jennifer Petticord, ELP/ELD Teacher/Specialist, Westlake City Schools, Ohio
- Janelle Poulson, Dual Language Teacher, Sioux City Community School District, Iowa
- Darla Proppe, District ESL Coordinator, Lake Hamilton School District, Arkansas
- Jennifer Prowell, ELP/ELD Teacher/Specialist, Centennial School District, Oregon
- Carol Richardson, ELL Department Head, Olympia School District, Washington
- Virgil Ruiz, Spanish Literacy/Bilingual Education Teacher, Forest Grove School District, Oregon
- Alice Shaffer, General Education Teacher, Manhattan-Ogden School District, Kansas
- Julia Shoemaker, ELL Director, Lonoke School District, Arkansas
- Vitaliy Shyyan, Research Associate, National Center on Educational Outcomes, Minnesota
- Daniel Spatzierath, ELP/ELD Teacher/Specialist, Mitchell Public Schools, Nebraska
- Paul Stieber, ELP/ELD Teacher/Specialist, Bethel School District, Oregon
- Dianne Sweeney, ELP/ELD Teacher/Reading Specialist, Shawnee Mission School District, Kansas
- Crystal Tate, ELP/ELD Teacher/Specialist, West Des Moines School District, Iowa

- Liz Toomey, ELP/ELD Teacher/Specialist, Dayton Public Schools, Ohio
- Kristin Tregillus, ELP/ELD Teacher/Specialist, Tukwila School District, Washington
- Sarah Wait, ELP/ELD Teacher/Specialist, Cedar Falls School District, Iowa
- Anita Wicker, ESL Coordinator, Nettleton School District, Arkansas
- Janet Wolf, Kindergarten Dual Immersion Teacher, Greater Albany Public Schools, Oregon
- Barbara Wright, ELP/ELD Teacher/Specialist, Topeka Public Schools, Kansas
- Valerie Zahuranec, ELP/ELD Teacher/Specialist, Greenbrier County School District, West Virginia

**Table 17: Demographic Information and Expertise of Educators Who Conducted Content Reviews**

Category	Responses	Number
State	Arkansas	6
	Florida	0
	Iowa	7
	Kansas	10
	Louisiana	0
	Minnesota	2
	Nebraska	9
	Ohio	4
	Oregon	12
	South Carolina	0
	Washington	9
	West Virginia	2
Gender	Female	52
	Male	9
Ethnicity	Asian/Pacific Islander	2
	Hispanic	7
	Other Hispanic/American Indian	0
	White	45
	Multi-Racial	0
	Prefer Not to Respond	7
Type of School District	Rural	24
	Urban	37
Current Position	Classroom/Content Teacher or General Education Teacher	4
	ELD/ELP Teacher Specialist	38
	District Administrator	4
	District ELL Coordinator	0
	District ESL Coordinator	1
	ESL Teacher Trainer	0
	ELL Instructional Coach	1
	ESL Assessment Coordinator	1
	ELL Assessment Coordinator	1
Higher Education	2	

	Building Administrator	1
	Unknown	8
Years of ELP/ELD teaching and leadership experience	Less than 1 year	1
	1 to 5 years	7
	6 to 10 years	21
	10 to 15 years	9
	16 to 20 years	7
	21 to 25 years	4
	25 years or more	2
	Unreported	10
Highest Level of Education	Bachelor's degree	7
	Master's degree	42
	Doctoral degree	6
	Unreported	6

**Educators Who Participated in Bias Review Committees**

- Sherry Bergman, ELP/ELD Teacher/Specialist, Clarkston School District, Washington
- Kimberly Berman, ELP/ELD Teacher/Specialist, Seattle Public Schools, Washington
- Renée Bohaty, ELP/ELD Teacher/Specialist, Lincoln Public Schools, Nebraska
- Penny Businga, ELP/ELD Teacher/Manager, Educational Service Unit #13, Nebraska
- Taffy Carlisle, ELD Program Assistant, Spokane Public Schools, Washington
- Amy Davis, ELP/ELD Teacher/Specialist, Wichita Public Schools, Kansas
- Carly Groszhan, ELP/ELD Teacher/Specialist, Seattle Public Schools, Washington
- Kelly Kitterman, District ELL Instructional Coordinator, Derby Public Schools USD 260, Kansas
- Amanda Levos, Coordinator for ELL and Migrant Education, Grand Island Public Schools, Nebraska
- Justin Luttrell, Content Teacher, Riverview School District, Arkansas
- Sara McColloch, Classroom Teacher, Des Moines Dioceses, Iowa
- Norma Mondragon, Content Teacher, Hamburg School District, Arkansas
- Joanie Monroy, Associate Professor of ESL/Bilingual Education, Heritage University, Washington
- Sidni Musick, ESL Adjunct Instructor, Garden City Community College, Kansas
- Tatiana Sildus, Associate Professor of TESOL, Pittsburg State University, Kansas
- Allison Smith, ELP/ELD Teacher/Specialist, Portland Public Schools, Oregon
- H.L. Smith-Pokrandt, ELP/ELD Teacher/Specialist, Rogers Public Schools, Arkansas
- Valerie Snyder, ELP/ELD Teacher/Specialist, Wichita Public Schools, Kansas
- Carrie Ann Tkaczyk, ELP/ELD Teacher/Specialist, North Clackamas School District, Oregon
- Ann Walker, ELP/ELD Teacher/Specialist, Wichita Public Schools, Kansas
- Grace Waylen, ELP/ELD Teacher/Specialist, Shoreline Public Schools, Washington

**Table 18: Demographic Information and Expertise of Educators Who Participated in Bias Review Committees**

Category	Responses	Number
State	Arkansas	3
	Florida	0
	Iowa	1
	Kansas	6
	Louisiana	0
	Nebraska	3
	Ohio	0
	Oregon	2
	South Carolina	6
	Washington	0
	West Virginia	0
Gender	Female	20
	Male	1
Ethnicity	Asian/Pacific Islander	0
	Hispanic	1
	Other Hispanic/American Indian	0
	White	18
	Multi-Racial	0
	Prefer Not to Respond	2
Type of School District	Rural	7
	Urban	14
Current Position	Classroom/Content Teacher or General Education Teacher	3
	ELD/ELP Teacher Specialist	13
	District Administrator	2
	District ELL Coordinator	1
	ESL Teacher Trainer	0
	ESL Assessment Coordinator	0
	Higher Education	2
Years of ELP/ELD teaching and leadership experience	Less than 1 year	0
	1 to 5 years	6
	6 to 10 years	2
	10 to 15 years	3
	16 to 20 years	3
	21 to 25 years	1
	25 years or more	1
	Unreported	5
Highest Level of Education	Bachelor's degree	2
	Master's degree	16
	Doctoral degree	3
	Unreported	0

## Appendix C: Sources of Data and Quality Control Process Implemented for This Report

The great majority of the data in this report are taken directly from the metadata files (described in Section 3.2.5.B) that are a key element of the ELPA21 item pool. As part of the production of this report, a significant effort was undertaken to quality check all of these metadata.

These steps were followed for that quality check of metadata:

- Each item included in the exported item pool was identified.
- A number of consistency checks were made on the metadata in the export files (for example, ensuring that metadata coded for match to standard and match to PLD were feasible, ensuring that coding of experimental items and of text complexity scores were consistent across sets).
- Fields that were known to have been in flux relatively late in the development process were checked at the level of individual values (for example, coding of items intended as experimental items, as discussed in Section 3.2.2).
- Data from the IBIS data bank was re-run and checked for consistency against the export files
- Any errors found (including some errors in text complexity coding, in coding of accessibility status, and missing “do not use with” information for some items) were corrected.
- Consistency checks were then re-run to ensure all data were correct before being re-exported.

The data tables in this report are based on the updated metadata files generated by the process described above. In order to produce the data tables, the following steps were taken:

- All metadata files were double checked to ensure that the source of the items was correct
- A master list for the data tables in the report that defines the process for generating each table was created and includes:
  - A definition of the source data;
  - Whether experimental items were to be included;
  - Whether set leaders were included;
  - Which data should be pivoted;
  - Which data should be included in totals.
- Pivot tables were built from the identified data.
- Each table was quality checked from the source and pivot tables to make sure that they matched the definition.
- Each table was cross checked against the other report tables to ensure consistency.

The report also contains data reporting on in-process steps of the development of the item pool that are not reflected in the item pool metadata. (For example, Table 6 reports on the state screening and educator review of reading passages.) These data were quality checked by reviewing against the original source documents and then reviewing for internal consistency.