



The ELPA21 Framework: Summative Assessment

School Year 2015 - 2016



TABLE OF CONTENTS

1. INTRODUCTION 1

 1.1. ELPA21 OVERVIEW 1

 1.2. ELPA21 BACKGROUND..... 1

 1.3. PURPOSE AND SCOPE OF THE ASSESSMENT FRAMEWORK 3

 1.4. RELATED REPORTS AND DOCUMENTS..... 3

2. ELP STANDARDS (COMPLETE) 3

3. ELPA21 THEORY OF ACTION..... 4

4. CLAIMS AND SUBCLAIMS 4

 4.1. CLAIMS 5

 4.2. SUBCLAIMS 5

5. OPERATIONAL BLUEPRINT 7

 5.1. GENERAL PRINCIPLES FOR TEST DEVELOPMENT 14

6. ACCESSIBILITY PRINCIPLES..... 14

 6.1. ACCESSIBILITY AND ACCOMMODATIONS MANUAL..... 15

 6.2. ACCOMMODATED FORMS – B/LV AND P&P..... 15

7. METADATA 16

 7.1. METADATA FOR FIELD TESTING 17

 7.2. METADATA FOR OPERATIONAL TESTING 17

8. SCORING RULES AND RUBRICS 17

 8.1. SCORING RULES FOR CR ITEMS 17

 8.2. SCORING RUBRICS FOR CR ITEMS 18

 8.3. SCORING RULES FOR TEIS 18

9. ITEM DEVELOPMENT PROCESS 19

 9.1. OPERATIONAL ITEM DEVELOPMENT 21

 9.2. OPERATIONAL ITEMS ALIGNED TO EACH PRACTICE 24

10. PROFICIENCY LEVELS FOR SCORING AND REPORTING 26

APPENDIX A: METADATA AND ANSWER KEY INTERPRETATION GUIDE..... 27

APPENDIX B: ELPA21 THEORY OF ACTION..... 31

APPENDIX C: PARTIAL CREDIT SCORING RULES VALIDATION REPORT..... 45

LIST OF TABLES

Table 5.1. ELPA21 Operational Summative Assessment Test Blueprints by Grade Band and Domain ...	9
Table 5.2. ELPA21 Tasks by Grade Band and Domain	10
Table 5.3. Measures of ELPA21 Standards	11
Table 9.1. Operational Items by Task and Grade Band	21
Table 9.2. Operational Items Aligned to Each ELP Standard	23
Table 9.3. Operational Items Aligned to Each Practice	25

1. INTRODUCTION

1.1. ELPA21 OVERVIEW

The English Language Proficiency Assessment for the 21st Century (ELPA21) is designed to measure the performance of English language learners (ELLs) as they progress through their K–12 public education and achieve college and career readiness. The assessment system is being developed by a consortium of states and is federally funded through September 2016 with a \$9.1 million grant from the U.S. Department of Education.

States participating in ELPA21 represent multiple regions of the United States and are politically and demographically diverse. There is also variety in how students are assessed in academic content in participating states: Some ELPA21 states have state-developed assessments and others participate in the Partnership for Assessment of Readiness for College and Careers (PARCC) or the Smarter Balanced Assessment Consortium. ELPA21 is collaborating with partners from the Understanding Language Initiative of Stanford University; the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) of the University of California, Los Angeles; the National Center on Educational Outcomes of the University of Minnesota (NCEO); and the Council of Chief State School Officers (CCSSO) to develop the assessment system. The Oregon Department of Education (ODE) is the lead state agency, and CCSSO is the project management partner. ELPA21 states will use the new ELPA21 assessment system beginning in the 2015–2016 school year.

The mission and vision of ELPA21 are as follows:

- **Mission:** Acknowledging the diverse and rich language experiences English language learners (ELLs) bring to school, we recognize their English language proficiency is constantly growing. ELPA21 measures that growth based on the new English Language Proficiency Standards and provides valuable information that informs instruction and facilitates academic English proficiency so that all ELLs leave high school prepared for college and career success.
- **Vision:** To provide assessments that best measure English language learners' mastery of the communication demands of states' rigorous academic standards.

1.2. ELPA21 BACKGROUND

ELPA21 is unique in many ways, starting with its standards. The new ELP Standards were developed by educators and state experts, the Understanding Language Initiative of Stanford University's Graduate School of Education, WestEd's Assessment and Standards Development Services, and CCSSO. The standards represent a significant shift for most states in what we now know about how ELLs learn English.

As ELLs practice and learn English in the classroom, they simultaneously interact with grade-level academic content. Increasing the expectations for the academic content that students must master in school requires a parallel increase in expectations for English language acquisition. The ELP Standards

describe these higher expectations by integrating language development with appropriate mathematics, language arts, and science practices by grade. The standards describe how language is used to meet the rigorous content demands in each grade and how students progress, by grade and grade band, toward English language proficiency. This is a very helpful tool for educators—both for those educating ELLs and for content area teachers. As ELLs learn the academic uses of the English language, they acquire the content knowledge necessary to be on track for college and career readiness.

Key questions ELPA21 states and collaborators asked to ensure the development of the best assessment system to meet the needs of ELLs in their states include:

- How do we move from the vision to the implementation of better standards and assessments for ELLs?
- What steps need to be taken to improve English language teaching and learning?
- What evidence exists that the identified steps and strategies will be successful in the context of ELPA21?
- As a result of our actions, what can we expect to happen?
- How will these actions impact students, educators, and schools?
- What results do we expect to see, and how will we measure them?
- What are indicators of success?

The ELPA21 system supports ELLs by determining initial proficiency and providing information to inform placement through a screener. The summative assessment informs decisions about student reclassification or continued placement; provides information that can help guide instruction and nurture student growths; and contributes accountability information for the overall system and each member state. ELPA21 uses an online testing platform to assess students' English language proficiency levels and progress in four domains: Reading, Writing, Listening, and Speaking. There are six ELPA21 grade bands: K, 1, 2–3, 4–5, 6–8, and 9–12.

Both the objectives of the assessment system and the standards underlying ELPA21 (CCSSO et al., 2014)¹, require the use of context-rich tasks that reflect, to the extent possible, the actual tasks students engage in as they learn academic content. The ELPA21 standards and assessment “[reframe] language proficiency from ‘*What language does the student have?*’ to ‘*What is the student able to do with language in the content areas?*’” (ELPA21, 2014, p. 7).²

¹ Council of Chief State School Officers (CCSSO), WestEd, & the Understanding Language Initiative at Stanford University. (2014). English language proficiency (ELP) standards. Retrieved from http://www.elpa21.org/sites/default/files/Final%204_30%20ELPA21%20Standards_1.pdf.

² English Language Proficiency Assessment for the 21st Century (ELPA21). (2014). *Theory of action*. Unpublished manuscript.

1.3. PURPOSE AND SCOPE OF THE ASSESSMENT FRAMEWORK

This framework provides an overview of the ELPA21 assessment system. The audience for this document is the assessment system user, including state-, district-, building- and classroom-level educators, researchers, psychometricians, and technical advisors. This framework contains the information and guidance needed to understand the theory and constructs underpinning ELPA21 and to develop items and operational test forms through its hyperlinks to critical documents. It describes the intended uses of the score data so that the assessment and results are helpful to those making important decisions about ELLs' learning. The framework is intended to be understandable and simple to navigate, and is structured for ease of access.

1.4. RELATED REPORTS AND DOCUMENTS

- *Appendix B: ELPA21 Theory of Action*
- *Appendix C: ELPA21 Partial Credit Scoring Rules Validation Report (ETS)*
- *Item Development Process Report* from the Educational Testing Service (ETS)
- *ELPA21 Spring 2015 Field Test Technical Report* from Questar Assessment, Inc. (Questar)

2. ELP STANDARDS (COMPLETE)

In 2013, CCSSO contracted with WestEd's Assessment and Standards Development Services to lead the development of a set of standards for English language proficiency that would correspond to college- and career-readiness standards for English language arts, mathematics, and science. The resulting English Language Proficiency (ELP) Standards correspond to states' rigorous content standards in English language arts, mathematics, and science. Beyond understanding common English usage, ELLs need to understand language used for grade-level instruction in English language arts, mathematics, and science.

In designing these standards, WestEd had these three goals:

1. The standards should correspond to, and be used in tandem with, the college and career readiness (CCR) standards for English language arts, mathematics, and science.
2. The standards should highlight and amplify the critical language, knowledge about language, and skills using language in CCR standards necessary for ELLs to be successful in school.
3. The standards should be simple and clear and should aim high so that teachers can focus on what is most important for college and career readiness.

The ELP Standards are understandable, usable, and easily transferable to classroom curricula and instruction for English language development. They are meaningful, coherent and rigorous, concise and measurable, and are both vertically and horizontally aligned. In order to create the ELP Standards, [WestEd](#) did the following:

- Developed proficiency-level descriptors

- Refined initial drafts, incorporated state feedback, and improved connections across Parts I, II, and III in spring 2013
- Analyzed how the ELP Standards correspond to the CCR standards

The complete ELP Standards are available on the following website:

<http://www.elpa21.org/elp-standards>

3. ELPA21 THEORY OF ACTION

Appendix B contains the *ELPA21 Theory of Action* document that describes how ELPA21 will facilitate the move from current English language expectations and instructional practice to those necessary for ELLs to gain proficiency in the academic language used within math, science, and ELA, and to ultimately become college- and career-ready. The *ELPA21 Theory of Action* describes how ELPA21's mission will be carried out, aligning intended assumptions with the organizational context of ELPA21. It connects strategy to action and identifies the multiple dependencies required for the successful implementation of the vision of ELPA21.

The *ELPA21 Theory of Action* is grounded in research and evidence-based practice and describes what the consortium intends to achieve. It also connects strategy and actions to objectives and desired outcomes to fulfill the mission and vision of ELPA21.

4. CLAIMS AND SUBCLAIMS

This section is excerpted from the *Item Development Process Report* from Educational Testing Service (ETS).³

One of the guiding principles for the development of the ELPA21 assessment system was to employ an evidence-centered design (ECD) approach to identify key claims and subclaims in the standards and to use those claims and subclaims to inform item development. 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).⁴ 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.

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 assessment system, it was necessary to articulate a principled manner of interpreting and sampling this proficiency so it could be measured within the confines of a standardized assessment with practical time limits. Although several of the

³ Hauck, M. C., Pooler, E., and Anderson, D. P. (2015). *ELPA21 item development process report*. Report submitted by Educational Testing Service (ETS), May 15, 2015.

⁴ 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.

ELPA21 ASSESSMENT FRAMEWORK

standards call for the integration of skills, Title III mandates that students be assessed in the four separate domains of Reading, Writing, Listening, and Speaking. While recognizing the multidimensional nature of English language development and the emphasis on collaborative skills in the ELP Standards, ELPA21 relies on test items designed to measure skills by the four domains.

4.1. CLAIMS

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

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

4.2. SUBCLAIMS

The ELPA21 subclaims represent a disaggregation of the 10 ELP Standards across the domains of Reading, Writing, Listening, and Speaking. The subclaim number refers to the ELP Standard from which the subclaim was derived. For example, 2W is the writing subclaim derived from Standard 2. Because not all claims are relevant to all of the four domains (e.g., Standard 3, which focuses on productive skills, maps on to subclaims for Speaking and Writing but not the receptive skills of Listening or Reading), there are a total of 26 subclaims.

Subclaim	In grade-appropriate contexts...
2W	The English language learner participates in written exchanges of information, ideas, and analyses, responding to peer, audience, or reader comments and questions.
3W	The English language learner writes about complex literary and informational texts and topics.
4W	The English language learner constructs written claims and supports them with reasoning and evidence.
5W	The English language learner conducts research, evaluating and communicating in writing, findings to answer questions or solve problems.
6W	The English language learner uses writing to analyze and critique arguments of others.
7W	The English language learner adapts language choices to purpose, task, and audience when communicating in writing.

Subclaim	In grade-appropriate contexts...
9W	The English language learner writes clear and coherent text.
10W	The English language learner uses standard English accurately to communicate in writing.
2S	The English language learner participates in spoken exchanges of information, ideas, and analyses, by orally responding to peer, audience, or reader comments and questions.
3S	The English language learner speaks about complex literary and informational texts and topics.
4S	The English language learner constructs spoken claims and supports them with reasoning and evidence.
5S	The English language learner conducts research, evaluating and orally communicating, findings to answer questions or solve problems.
6S	The English language learner uses oral language to analyze and critique arguments of others.
7S	The English language learner adapts language choices to purpose, task, and audience when speaking.
9S	The English language learner expresses her/himself clearly and coherently in oral communication.
10S	The English language learner uses standard English when communicating orally.
1R	The English language learner constructs meaning from literary and informational text.
2R	The English language learner demonstrates comprehension of written exchanges of information, ideas, and analyses.
5R	The English language learner conducts research based on written sources of information and demonstrates comprehension by evaluating written findings.
6R	The English language learner analyzes and critiques arguments of others that are presented in writing.
8R	The English language learner determines the meaning of words and phrases in literary and informational text.
1L	The English language learner constructs meaning from oral presentations and literary and informational text.
2L	The English language learner demonstrates comprehension of oral exchanges of information, ideas, and analyses.
5L	The English language learner conducts research and demonstrates comprehension by evaluating findings presented orally.
6L	The English language learner analyzes and critiques the oral arguments of others.

Subclaim	In grade-appropriate contexts...
8L	The English language learner determines the meaning of words and phrases in oral presentations.

5. OPERATIONAL BLUEPRINT

The ELPA21 test blueprints are documents, organized by domain, that define what each test form will contain. These blueprints serve as guiding documents for 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 serve as a guide for how to populate the field test pool with enough coverage to allow for full operational forms and a screener.⁵

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 detailed the number of items to be included on test forms at each grade or grade band as distributed across task types and response formats: selected response (SR), technology enhanced (TE), short constructed response (CR), and extended (CR). The test blueprints also showed how many total items each test form would contain and an estimate of how many score points would be generated for each grade or grade band in each domain. The initial drafts of these test blueprints were reviewed extensively by the Item Acquisition and Development (IAD) and Assessment Design and Scaling (ADS) Task Management Teams (TMTs). A number of adjustments were made 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.

Stage 2 test blueprints served as a more detailed plan for the development of the operational summative and screener assessments. The operational summative assessment blueprints reflect the numbers of score points associated with CR and TE tasks after their rubrics and scoring rules have been validated based on field test results and reviewed by the IAD TMT to ensure that the assessment appropriately measures the domains of Reading, Writing, Listening, and Speaking.

The summative blueprints are organized by domain and task, and ensure an appropriate number of score points are attributed for each domain while ensuring total testing time is appropriate for the students in that grade level. A task describes the type of interaction elicited from the students by either a discrete item or by a set of items.

Table 5.1 provides the operational test blueprints for each grade level and domain, and the number of tasks and points for each domain. Table 5.2 provides a summary of the tasks by domain for each grade

⁵ Hauck, M. C., Pooler, E., and Anderson, D. P. (2015). *ELPA21 item development process report*. Report submitted by Educational Testing Service (ETS), May 15, 2015.

 **ELPA21 ASSESSMENT FRAMEWORK**

level that will appear on each summative assessment, and Table 5.3 provides the tasks used to measure each ELPA21 standard.

Table 5.1. ELPA21 Operational Summative Assessment Test Blueprints by Grade Band and Domain

Grade Band	Domain	#Tasks	#Points
K	Listening	15	28
	Reading	13	23
	Speaking	6	27
	Writing	13	21
1	Listening	14	24
	Reading	21	29
	Speaking	5	25
	Writing	14	21
2–3	Listening	14	24
	Reading	16	34
	Speaking	5	25
	Writing	14	24
4–5	Listening	16	33
	Reading	11	28
	Speaking	5	30
	Writing	11	30
6–8	Listening	19	34
	Reading	8	33
	Speaking	4	27
	Writing	6	28
9–12	Listening	12	26
	Reading	9	35
	Speaking	4	27
	Writing	6	28

Table 5.2. ELPA21 Tasks by Grade Band and Domain

Grade Band	Listening	Reading	Speaking	Writing
K	<ul style="list-style-type: none"> • Follow Instructions • Listen and Match <ul style="list-style-type: none"> - Phrase - Sentence - Word • Long Conversation • Read-Aloud Story • Short Conversation • Teacher Presentation 	<ul style="list-style-type: none"> • Informational Set • Read and Match <ul style="list-style-type: none"> - Phrase - Sentence - Word • Read-Along Story • Short Correspondence • Word Wall 	<ul style="list-style-type: none"> • Classroom Tableau • Observe and Report • Picture Description • Show and Share Presentation • Show and Share Questions 	<ul style="list-style-type: none"> • Complete the Story • Sentence Builder • Word Builder <ul style="list-style-type: none"> - Phrase - Sentence - Word • Paper and Pencil <ul style="list-style-type: none"> - Complete a Word - Copy a Word - Opinion - Write a Sentence - Write a Word
1	<ul style="list-style-type: none"> • Follow Instructions • Listen and Match <ul style="list-style-type: none"> - Sentence - Word • Long Conversation • Read-Aloud Story • Short Conversation • Teacher Presentation 	<ul style="list-style-type: none"> • Informational Set • Literary Set • Procedural Text • Read and Match <ul style="list-style-type: none"> - Sentence - Word • Read for Details • Read-Along Sentence • Short Correspondence 	<ul style="list-style-type: none"> • Classroom Tableau • Conversation • Observe and Report • Opinion • Picture Description 	<ul style="list-style-type: none"> • Sentence Builder • Word Builder • Paper and Pencil <ul style="list-style-type: none"> - Copy a Word - Storyboard - Write a Sentence - Write a Word
2–3	<ul style="list-style-type: none"> • Follow Instructions • Listen and Match <ul style="list-style-type: none"> - Sentence - Word • Long Conversation • Read-Aloud Story • Short Conversation • Teacher Presentation 	<ul style="list-style-type: none"> • Informational Set • Literary Set • Procedural Text • Read and Match <ul style="list-style-type: none"> - Sentence - Word • Read for Details • Read-Along Sentence • Short Correspondence 	<ul style="list-style-type: none"> • Classroom Tableau • Compare Pictures • Conversation • Observe and Report • Opinion 	<ul style="list-style-type: none"> • Opinion • Picture Caption • Sentence Builder • Storyboard • Word Builder
4–5	<ul style="list-style-type: none"> • Follow Instructions • Interactive Student Presentation • Listen and Match <ul style="list-style-type: none"> - Sentence - Word • Listen for Information • Short Conversation • Student Discussion • Teacher Presentation: Read Aloud 	<ul style="list-style-type: none"> • Extended Informational Set • Extended Literary Set • Match Picture to Word and Sentence • Short Correspondence Set • Short Informational Set • Short Literary Set 	<ul style="list-style-type: none"> • Analyze a Visual • Compare Pictures • Conversation • Language Arts Presentation • Observe and Report 	<ul style="list-style-type: none"> • Discrete Editing Tasks • Sentence Builder • Storyboard • Word Builder • Write an Opinion • Writing Questions Task

Grade Band	Listening	Reading	Speaking	Writing
6–8	<ul style="list-style-type: none"> • Academic Debate • Academic Lecture or Discussion • Follow Instructions • Interactive Student Presentation • Listen and Match <ul style="list-style-type: none"> – Sentence – Word • Listen for Information • Short Conversation 	<ul style="list-style-type: none"> • Argument and Support Essay Set • Extended Informational Set • Extended Literary Set • Short Informational Set • Short Literary Set • Short Paragraph 	<ul style="list-style-type: none"> • Analyze a Visual and a Claim • Compare Pictures • Language Arts Presentation • Observe and Report 	<ul style="list-style-type: none"> • Construct a Claim • Discrete Editing Tasks • Respond to a Peer E-mail • Storyboard • Writing Questions Task
9–12	<ul style="list-style-type: none"> • Academic Debate • Academic Lecture and Discussion • Interactive Student Presentation • Listen and Match <ul style="list-style-type: none"> – Sentence – Word • Listen for Information • Short Conversation 	<ul style="list-style-type: none"> • Argument and Support Essay Set • Discrete Items • Extended Informational Set • Extended Literary Set • Short Informational Set • Short Literary Set 	<ul style="list-style-type: none"> • Analyze a Visual and a Claim Argument • Compare Pictures • Language Arts Presentation • Observe and Report 	<ul style="list-style-type: none"> • Construct a Claim • Discrete Editing Tasks • Respond to a Peer E-mail • Storyboard • Writing Questions Task

Table 5.3. Measures of ELPA21 Standards

Tasks	1	2	3	4	5	6	7	8	9	10
Listening										
Follow Instructions	K, 1, 2-3, 4-5, 6-8	4-5						K, 1, 2-3, 4-5, 6-8		
Listen and Match	K, 1, 2-3, 4-5, 6-8, 9-12							K, 1, 2-3, 4-5, 6-8, 9-12		
Listen for Information	4-5, 6-8, 9-12				4-5, 6-8, 9-12			4-5, 9-12		
Read-Aloud Story	K, 1, 2-3							K, 1, 2-3		
Short Conversation	1, 4-5, 6-8, 9-12	K, 1, 2-3, 4-5, 6-8, 9-12				4-5, 9-12		K, 1, 2-3, 9-12		
Long Conversation	1	K, 1, 2-3				2-3		K		
Student Discussion	4-5	4-5			4-5	4-5				
Academic Debate	6-8	9-12				6-8, 9-12		9-12		


ELPA21 ASSESSMENT FRAMEWORK

Tasks	1	2	3	4	5	6	7	8	9	10
Academic Lecture or Discussion	6-8, 9-12							6-8, 9-12		
Teacher Presentation	K, 1, 2-3, 4-5	1			K, 2-3	4-5		K, 1, 2-3		
Interactive Student Presentation	4-5, 6-8, 9-12	4-5, 9-12			4-5, 6-8	9-12		9-12		
Reading										
Informational Set	K, 1, 2-3							K, 1, 2-3		
Extended Informational Set	4-5, 6-8, 9-12	4-5, 6-8, 9-12			4-5, 6-8, 9-12	4-5, 9-12		4-5, 6-8, 9-12		
Short Informational Set	4-5, 6-8, 9-12	4-5, 6-8, 9-12			4-5, 6-8, 9-12	6-8, 9-12		6-8, 9-12		
Literary Set	1, 2-3				4-5			1, 2-3		
Extended Literary Set	4-5, 6-8, 9-12	4-5, 6-8				4-5, 6-8		4-5, 6-8, 9-12		
Short Literary Set	4-5, 6-8, 9-12	4-5, 6-8				4-5		4-5, 6-8, 9-12		
Procedural Text	1, 2-3							1, 2-3		
Read and Match	K, 1, 2-3							K, 1, 2-3		
Read for Details	1, 2-3							1, 2-3		
Read-Along Story	K							K		
Read-Along Sentence	1, 2-3							2-3		
Match Picture to Word and Sentence	4-5							4-5		
Short Correspondence	K, 1, 2-3, 4-5	K, 1, 2-3, 4-5			K			K, 1, 2-3		
Short Correspondence Set						4-5		4-5		
Word Wall	K				K			K		
Short Paragraph	6-8	6-8						6-8		
Argument and Support Essay Set	6-8, 9-12	6-8, 9-12			6-8, 9-12	6-8, 9-12		6-8, 9-12		
Discrete items	9-12							9-12		
Speaking										
Classroom Tableau			K, 1, 2-3							K, 1, 2-3
Conversation		1, 2-3	1						4-5	1, 4-5
Analyze a Visual		4-5	4-5	4-5					4-5	4-5
Analyze a Visual and a Claim		6-8	6-8, 9-12	6-8, 9-12	6-8, 9-12	6-8, 9-12			6-8	6-8


ELPA21 ASSESSMENT FRAMEWORK

Tasks	1	2	3	4	5	6	7	8	9	10
Compare Pictures		6-8	2-3, 4-5, 6-8, 9-12						4-5, 6-8	2-3, 4-5, 6-8
Observe and Report		6-8	K, 1, 2-3, 4-5, 6-8, 9-12		K, 1, 6-8		4-5, 6-8, 9-12		K, 1, 2-3, 4-5, 6-8, 9-12	K, 1, 4-5, 6-8, 9-12
Opinion		1	1	1, 2-3						1
Picture Description			K, 1							K, 1
Conversation		4-5	4-5							
Show and Share Presentation		K		K						K
Show and Share Questions		K								K
Language Arts Presentation		6-8, 9-12	4-5, 6-8, 9-12	4-5, 6-8, 9-12	6-8	6-8			4-5, 6-8	4-5, 6-8
Writing										
PP-Copy a Word			K, 1							K, 1
PP-Opinion			K	K						K
PP-Storyboard			1						1	1
PP-Write a Sentence			K, 1						1	K, 1
PP-Write a Word			K, 1						1	K, 1
Complete the Story			K							K
Discrete editing tasks									9-12	4-5, 6-8, 9-12
Sentence Builder			K, 1, 2-3, 4-5						1, 4-5	K, 1, 2-3, 4-5
Word Builder			K, 1, 2-3, 4-5						4-5	K, 2-3, 4-5
Opinion		2-3		2-3						
Write an opinion		4-5	4-5	4-5			4-5		4-5	4-5
Picture Caption			2-3							2-3
Storyboard		6-8	2-3, 4-5, 6-8, 9-12				4-5, 6-8, 9-12		2-3, 4-5, 9-12	4-5, 6-8, 9-12
Writing questions task		4-5, 6-8, 9-12					4-5, 6-8, 9-12		4-5, 9-12	4-5, 6-8, 9-12
Construct a claim		6-8	6-8	6-8, 9-12			6-8, 9-12		6-8, 9-12	6-8, 9-12
Respond to a Peer E-mail		6-8, 9-12	6-8	6-8, 9-12					9-12	6-8, 9-12

5.1. GENERAL PRINCIPLES FOR TEST DEVELOPMENT

Test Development for the ELPA21 assessment program includes developing test forms, conducting quality assurance over all testing materials, and providing high-quality editorial review and proofing of the test forms, items that make up the item bank, and final assessments.

In December 2014, Questar Assessments facilitated item reviews by the ELPA21 consortium leading up to the field test assessments. The field test forms were developed in collaboration with the ELPA21 IAD TMT, using a block design to ensure all items were administered in the field tests that began in February 2015.

The field test window was open February 2 to March 31. Following the field test window, Questar led an extensive data review with the ELPA21 consortium members and ELL educators in August 2015. The results of the data review were then implemented to determine the item pool available for operational use in the 2016 assessment year.

The ELPA21 TMT leads and CRESST analyzed the results from the field test data analyses and data review meetings against the operational blueprints to ensure the final operational item pool supported the test design. Initial test form builds began in August 2015. As the forms were built and reviewed by ELPA21 TMT leads, attention was focused on the test forms' validity and reliability in measuring the domains of Speaking, Listening, Reading, and Writing.

The final operational test forms were developed in September and October 2015, and released to operational vendors in November 2015.

During this time, test blueprints and forms were developed for the paper-pencil and the blind/low vision accommodated forms. The Questar ELL assessment specialists informed these blueprints to ensure the items used on the forms were appropriate for the administration format while ensuring the standards were appropriately and thoroughly assessed.

6. ACCESSIBILITY PRINCIPLES

Ensuring that ELPA21 test items are maximally accessible for as wide a range of students as is possible without compromising the measurement goals of the assessment was a high priority for the item design and development effort. Steps taken to enhance accessibility included:

- developing all art and graphics following best practices for accessibility
- developing all items to be compatible with the Accessible Portable Item Protocol (APIP)
- employing alt text processes on all items when possible without interfering with the target construct
- providing multiple presentations of test material (i.e., audio as well as written) when it is consistent with the construct of interest to do so

6.1. ACCESSIBILITY AND ACCOMMODATIONS MANUAL

The *ELPA21 Accessibility and Accommodations Manual* describes methods for making ELPA21 as accessible as possible, including a list of accessibility features available to all students on the administration platform, available to any student with teacher consent, and available to students with IEPs. Item writers were told to familiarize themselves with the accessibility tools and accommodations that would be available on ELPA21 and to consider these as they were developing items.

The Accessibility and Accommodations Manual was developed prior to the ELPA21 field test, and the field test environment mimicked an operational environment, allowing students to try out the assessment with their accommodations in place. The Accessibility and Accommodations Manual was then revised for operational implementation and for delivery to ELPA21 states, which will distribute the manual to their test administrators and educators.

6.2. ACCOMMODATED FORMS – B/LV AND P&P

Both of the accommodated forms for ELPA21 were developed to measure the same assessment targets as the main, online form so that all students assessed by ELPA21 would receive results that described their language proficiency according to the same set of standards.

The blind and low vision forms consist of tasks developed using the online tasks as models through a process called “twinning”. According to the ETS ELPA21 Item Development Process Report (pp 55–56), “twinning” is:

“... 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.”

These “twin” tasks use additional presentation and response modes that are appropriate for students who are blind or have low vision, such as a braille presentation of reading items, text-based stimuli, and response options and manipulative-based responses.

The paper and pencil forms consist of the online tasks converted to a format appropriate for a paper-based representation of the online item formats, either as is or via research-based format adjustments, such as displaying response options in matrix form. In cases where tasks from the online forms were deemed to be inappropriate for conversion to a paper and pencil format, the standards measured by these online technology-based items were examined and other tasks were considered for inclusion in the paper and pencil forms to ensure the standards were adequately represented in the paper and pencil test forms.

Additional details on the blueprints for the blind and low vision, paper, and large print forms will be available in version 2.0 of the ELPA21 Assessment Framework., to be released in early 2016.

7. METADATA

The following text is from the *ELPA21 Item Development Process Report* from ETS (pp. 27–28).⁶ See Appendix A for a table that defines the item metadata fields.

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 for 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.

⁶ Hauck, M. C., Pooler, E., and Anderson, D. P. (2015). *ELPA21 item development process report*. Report submitted by Educational Testing Service (ETS), May 15, 2015.

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.

7.1. METADATA FOR FIELD TESTING

Questar received metadata for field test ready items from ETS and used the metadata to build the field test forms beginning in November 2014. Throughout the build process, items were reviewed by the IAD TMT, as well as internally by Questar. As the Questar Technology Team implemented the items in field test forms, additional decisions were made that impacted some metadata fields. One example being the set leader behavior changed from “set leader above items” to “Set beside items” in order for the item to be presented appropriately to the students. These decisions were based on the original intent for how the item or task was to measure the construct and the decisions were made in consultation with the ELPA21 IAD TMT.

Throughout field testing and after, specific metadata were updated based on the findings from the field testing process, including max points, item enemy information, and answer keys for those items included in partial credit rule decisions. Answer keys were also corrected based on the results of key verification processes and decisions on how to represent the answer key (such as with partial credit rule decisions).

7.2. METADATA FOR OPERATIONAL TESTING

The final set of metadata used for operational testing represents the bulk of the metadata documented by ETS, as described in the introduction to this section 7, as well as the adjustments based on decisions made throughout the field testing process. The final set of metadata for operational testing were used to build operational, summative forms, and will be used to build future forms for the screener assessment and to support result-reporting functions.

8. SCORING RULES AND RUBRICS

8.1. SCORING RULES FOR CR ITEMS

All constructed response (CR) items from the Speaking and Writing domains are handscored by properly trained readers (scorers) using scoring rubrics provided by ELPA21. The scorers are trained on each task type using specific training materials, including scoring guides, training sets, and qualifying

sets. The scoring process is monitored by a scoring director for accuracy of scoring by using a read-behind process whereby a team leader or scoring director takes random student responses to check the accuracy of the scores or prescribed student responses if a prior scoring accuracy issue has been brought to the team leader or scoring director's attention.

For some tasks within the Speaking domain, items within a task set are scored as a cluster. In these cases, the task type has multiple tasks attached to a common set leader passage which require scoring all responses as a set and assigning one holistic score. These task types are contained in Grade/Grade Bands K through 4-5 for Show and Share Questions, Show and Share Presentation, Picture Description, Observe and Report, and Conversation.

8.2. SCORING RUBRICS FOR CR ITEMS

Items in the writing and speaking domains are scored against holistic rubrics for each task type.

The holistic scoring rubrics are the guide used by scorers (readers) during a human scoring process. The scorers are trained on the application of the scoring rubrics by first receiving training on the task types, practicing with a training set of items, and through discussions to ensure each scorer is consistently and accurately applying the rubric to student responses. Once the scorers are calibrated to apply the rubric consistently and reliably, the actual scoring process may begin.

At any time during the scoring process, scorers are directed to send up for review any student response that suggests the possibility of teacher interference, plagiarism, or use of inappropriate content. Similarly, scorers are instructed as to what kinds of response characteristics should trigger a review for disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). When a scorer identifies a response that fit these criteria, it is scored and then marked as an alert within the scoring system. The scorer selects the reason for the alert and includes any comments to explain the need for the alert.

All alerts go directly to a scoring director to ensure the responses are properly flagged. After the scoring director reviews all alerted responses, a file of alerted responses is generated. Alert files are sent to the ELPA21 team.

8.3. SCORING RULES FOR TEIS

The following text is derived from the *ELPA21 Partial Credit Scoring Rules Validation Report* from ETS (pp. 2–3).⁷ See Appendix C for the whole report.

In 2014, ETS, in collaboration with CCSSO and the ELPA21 Consortium, designed and developed a pool of test items based on the English Language Proficiency (ELP) Standards to form the basis of the

⁷ Pooler, E., Wang, J., and Doyle, B. (2015). *ELPA21 partial credit scoring rules validation report*. Report submitted by Educational Testing Service (ETS), September 9, 2015.

ELPA21 assessment system. The ELPA21 item pool contains a number of innovative task types, including several technology-enabled (TE) item types. As part of the item design and development effort, ETS developed preliminary scoring rules for these item types, including scoring rules for awarding partial credit where appropriate (Hauck et al., in press, pp. 25–26).

For ELPA21, TE items have been defined as “those computer-delivered items that include specialized interactions in the student response format or in the use of response data” (Hauck et al., in press, p. 25). Responding to a TE item requires a specialized interaction that can be more complex than responding to selected-response or text-entry (keyboarding/typing) items.

Because TE items can include more complex interactions, responses may provide more or different information about a student’s knowledge, skills, or abilities than a typical single-selection multiple-choice item. Whereas a typical single-selection multiple-choice item can only be scored dichotomously (correct or incorrect), some TE items can potentially be scored along a continuum of partially correct to fully correct depending on the number of interactions in an item. In addition to TE items, several multiple-select multiple-choice items for the ELPA21 assessment were also designed with the intent to apply partial credit scoring. For the purposes of ELPA21, partial credit scoring is defined as a proportion of credit awarded for a student response to a TE or multiple-select multiple-choice item that demonstrates some, but not total, success in completing the task.

The partial credit scoring rules for ELPA21 were not based on item complexity (e.g., the number of interactions required to respond to an item) or item difficulty. Instead, the partial credit scoring rules were related to the test construct and were based on distinctions among the ELPA21 proficiency-level descriptors (PLDs). That is, responses awarded a proportion of credit would provide evidence of students’ level of English language proficiency, as described in the PLDs.

Additionally, the scoring rules validation required that responses to proposed partial credit items have the following characteristics.

- The score obtained for each item should be positively related to student proficiency; that is, the student’s overall score should be positively correlated to the TE item score.
- The score levels described in the scoring rules should be supported by actual student responses; that is, the distinction between score levels should be clearly demonstrated by student responses, and each of the distinct score levels must be observed in the empirical data.

9. ITEM DEVELOPMENT PROCESS

ETS provided the *Item Development Process Report* on May 15, 2015.⁶ The document summarizes the activities undertaken by ETS in 2014 on behalf of and in collaboration with the CCSSO and the ELPA21 consortium related to the design and development of a pool of test items for the ELPA21 assessment system. The following is an overview of the content of the report and ETS’ role in item development.

Goals central to the 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.

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).⁸ 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, subclaims, 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.

The major deliverable produced at the end of the ETS’ item development work was the design and development of the field test pool for the ELPA21 assessment system, sufficient to support initial field testing that would lead to the development of initial operational forms of a screener and summative assessment.

⁸ 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.

The pool delivered for field testing contained 2,619 items, 2,469 of which were intended for initial operational use and 150 of which were 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 contained accessible content authored via APIP. Of the items intended for initial operational use, 1,138 or 46 percent were 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.

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

9.1. OPERATIONAL ITEM DEVELOPMENT

Questar received from ETS the field-test ready item pool, as described in the prior introduction, to assemble field test forms, administer the field test, then conduct data reviews in order to establish a final operational item pool.

The results of the data review included only 70 items sent for future revision and 8 items rejected. As a result the operational item pool is a robust pool that adequately supports the original vision for the ELPA21 assessment program.

Table 9.1 shows the final number of items for each task that appears in the summative ELPA21 assessments, and Table 9.2 shows the number of items aligned to each ELPA21 standard.

Table 9.1. Operational Items by Task and Grade Band

Domain	Task Type	Grade Band						Total
		K	1	2–3	4–5	6–8	9–12	
Listening	Academic Debate	--	--	--	--	17	15	32
	Academic Lecture and Discussion	--	--	--	--	--	18	18
	Academic Lecture or Discussion	--	--	--	--	8	--	8
	Follow Instructions	53	19	16	10	5	--	103
	Interactive Student Presentation	--	--	--	13	10	12	35
	Listen and Match	54	46	41	28	36	25	230

ELPA21 ASSESSMENT FRAMEWORK

Domain	Task Type	Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
	Listen for Information	--	--	--	12	11	15	38
	Long Conversation	17	12	12	--	--	--	41
	Read-Aloud Story	18	16	17	--	--	--	51
	Short Conversation	3	7	7	12	20	17	66
	Student Discussion	--	--	--	13	--	--	13
	Teacher Presentation	18	14	20	--	--	--	52
	Teacher Presentation: Read Aloud	--	--	--	15	--	--	15
	Listening Total	163	114	113	103	107	102	702
Reading	Argument and Support Essay Set	--	--	--	--	14	20	34
	Discrete Items	--	--	--	--	--	36	36
	Extended Informational Set	--	--	--	23	12	18	53
	Extended Literary Set	--	--	--	22	15	19	56
	Informational Set	15	23	22	--	--	--	60
	Literary Set	--	20	15	--	--	--	35
	Match Picture to Word and Sentence	--	--	--	25	--	--	25
	Procedural Text	--	19	15	--	--	--	34
	Read and Match	51	44	17	--	--	--	112
	Read for Details	--	4	5	--	--	--	9
	Read-Along Sentence	--	22	20	--	--	--	42
	Read-Along Story	18	--	--	--	--	--	18
	Short Correspondence	18	14	29	--	--	--	61
	Short Correspondence Set	--	--	--	16	--	--	16
	Short Informational Set	--	--	--	16	12	18	46
	Short Literary Set	--	--	--	11	17	16	44
	Short Paragraph	--	--	--	--	24	--	24
Word Wall	30	--	--	--	--	--	30	
Reading Total	132	146	123	113	94	127	735	
Speaking	Analyze a Visual	--	--	--	12	--	--	12
	Analyze a Visual and a Claim	--	--	--	--	10	--	10
	Analyze a Visual and a Claim Argument	--	--	--	--	--	10	10
	Classroom Tableau	36	30	25	--	--	--	91
	Compare Pictures	--	--	6	6	5	5	22
	Conversation	--	15	18	24	--	--	57
	Language Arts Presentation	--	--	--	18	15	15	48
	Observe and Report	16	4	3	3	3	3	32
	Opinion	--	12	6	--	--	--	18
	Oral Vocabulary	--	--	--	20	20	20	60

Domain	Task Type	Grade Band						Total
		K	1	2-3	4-5	6-8	9-12	
	Picture Description	30	6	--	--	--	--	36
	Show and Share Presentation	24	--	--	--	--	--	24
	Show and Share Questions	12	--	--	--	--	--	12
	Speaking Total	118	67	58	83	53	53	432
Writing	Complete the Story	12	--	--	--	--	--	12
	Construct a claim	--	--	--	--	3	4	7
	Discrete editing tasks	--	--	--	9	9	10	28
	Opinion	--	--	5	--	--	--	5
	Picture Caption	--	--	12	--	--	--	12
	Respond to a Peer E-mail	--	--	--	--	5	3	8
	Sentence Builder	12	24	30	19	--	--	85
	Storyboard	--	--	5	6	5	5	21
	Word Builder	36	29	16	8	--	--	89
	Write an Opinion	--	--	--	6	--	--	6
	Writing Questions Task	--	--	--	15	15	15	45
Writing Total	60	53	68	63	37	37	318	
Grand Total	473	380	362	362	291	319	2,187	

Table 9.2. Operational Items Aligned to Each ELP Standard

NOTE: Many items are aligned to multiple standards.

Grade Band	Domain	Standard									
		1	2	3	4	5	6	7	8	9	10
K	Listening	143	20	--	--	18	--	--	163	--	--
	Reading	115	18	--	--	48	--	--	132	--	--
	Speaking	--	36	82	24	16	--	--	--	16	118
	Writing			60			--	--	--	--	60
1	Listening	114	33	--	--	--	--	--	102	--	--
	Reading	146	14	--	--	--	--	--	102	--	--
	Speaking	--	27	67	12	4	--	--	--	4	67
	Writing	--	--	53	--	--	--	--	--	24	24
2-3	Listening	93	19	--	--	19	12	--	101	--	--
	Reading	111	29	--	--	--	--	--	123	--	--
	Speaking	--	18	34	6	--	--	--	--	3	31
	Writing	--	5	63	5	--	--	--	--	5	58
4-5	Listening	103	48	--	--	28	10	--	40	--	--
	Reading	113	52	--	--	19	13	--	44	--	--
	Speaking	--	30	63	12	--	--	3	--	77	83

Grade Band	Domain	Standard									
		1	2	3	4	5	6	7	8	9	10
	Writing	--	21	37	6	--	--	27	--	52	69
6-8	Listening	96	20	--	--	14	17	--	42	--	--
	Reading	84	80	--	--	11	13	--	34	--	--
	Speaking	--	53	53	18	14	11	3	--	64	67
	Writing	--	28	13	8	--	--	23	--	23	40
9-12	Listening	87	44	--	--	15	18	--	102	--	--
	Reading	127	55	--	--	47	24	--	127	--	--
	Speaking	--	15	53	14	10	5	3	--	3	3
	Writing	--	18	5	7	--	--	24	--	37	37

9.2. OPERATIONAL ITEMS ALIGNED TO EACH PRACTICE

Table 9.3 shows the number of operational ELPA21 items that were aligned to English language arts, mathematics, and science practices. Some items were aligned to multiple practices.

Table 9.3. Operational Items Aligned to Each Practice

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	
K	Listening	163			18			1					4		1						19	
	Reading	132			30		18														63	
	Speaking	102			16		82						37						83		167	
	Writing		60				3															
1	Listening	88					7	19					19			19	1				33	
	Reading	124					22						14				7	7			84	
	Speaking				4	27	67												4		44	
	Writing		53				48	24					43								48	
2-3	Listening	37				19	57														20	
	Reading	30			17	29	76														11	
	Speaking			6		18	31												3		3	
	Writing		22	5			46															
4-5	Listening	84		23	34		12	3		24			2	20		36	24		2		55	
	Reading	84		16	87	38	14	66	15	6		80		116		15	19	23	29		374	
	Speaking			12	39	36	83			36			12	30		7	3	3		66	161	
	Writing		63	6		27	58			12			11	55	30	42				24	92	
6-8	Listening	77		1	8	1		1		1			1			9	5		2	1	16	
	Reading	94		3					3								6	2			19	
	Speaking	50		9	4		47		2	16	1		19				29				42	
	Writing	32	18	9		16	22															
9-12	Listening	101		7	8	4	1			10	1		43							11	127	
	Reading	127		5	91	2	9	16		7			64		18		11			5	216	
	Speaking	29		18	10	14	9			14			46	3			9		38	28	110	
	Writing		29	4		4	33	15		8			50		30		8		16	23	74	

10. PROFICIENCY LEVELS FOR SCORING AND REPORTING

Questar will add the interim (range) PLD description once those are reviewed by ELPA21. Therefore, this section will be provided in Version 2.0 of this assessment framework.

APPENDIX A: METADATA AND ANSWER KEY INTERPRETATION GUIDE

Column	Column Header	Definition
A	Test	ELPA21
B	Administration	Operational (Summative) or Screener with the year noted (OP16)
C	Form	Test Form number (1, 2, 3,...N)
D	Position	Position in form (1, 2, 3,...N)
E	Task Set	Name or number for a task set (i.e., a set of items within a specific task)
F	Set Leader Accnum	8-digit code for associated set leader
G	Item Accnum	8-digit item code
H	Block	Test Block (applicable to ELPA21 Field Test only)
I	Seq Num	Sequence in block (applicable to ELPA21 Field Test only)
J	Slot	Slot position in block (applicable to ELPA21 Field Test only)
K	Set Leader Behavior	Expected rendering behavior of set leader
L	Item Level	Grade band of item
M	Domain	Content domain
N	Task Type	Item task type
O	Task Sub Type	Subtask type
P	Item Passage Seq	Item passage sequence
Q	Item Type	ELPA item type
R	Standard(s)	Pipe delimited list of associated ELPA standards
S	PLD(s)	Comma-delimited list of associated PLDs
T	Evidence Statement(s)	Pipe delimited list of associated evidence statements
U	ELPA21 Sub-Claim	Pipe delimited list of associated sub claims
V	Name	Item title
W	Answer Key Text	Comma-separated list of answer keys (see supplemental Answer Key Interpretation Guide below*)
X	Max Score Pts	Maximum score for the item
Y	Gradeband	K, 1, 2-3, 4-5, 6-8, or 9-12
Z	Academic Content Area	Content area of item content
AA	Modality	Item modality
AB	Text Complexity	Numerical text complexity of associated set leader
AC	Accessibility Concerns	Potential item level accessibility concerns

Column	Column Header	Definition
AD	Accessibility Review	Adaptation required to make the item accessible to populations
AE	Accessibility Feature	Alternate format for delivering item content to the candidate
AF	Alt text code	Category of alt text (1 - image not critical to content, 2 - image related to content but possibly still accessible, 3 - item is not accessible without image, 4 - not accessible)
AG	Technology Enabled	Identifies items with technology-related stimuli
AH	Technology Enhanced	Identifies items with technology-related response formats
AI	Technology Enhancement	Technology-related media enhancing item
AJ	Cluster Scoring	Set members to be scored as a combined score
AK	Experimental Item	Indicates an item is experimental integrated
AL	Item Enemy	Accnum of item(s) not to be included on same test form
AM	Source	Source of items – ETS or educator
AN	APIP	Indicator if item follows APIP standards
AO	Response Format	CR – constructed response, TE – technology enhanced, SR – selective response

*ELPA21 Answer Key Interpretation Guide

1. Multiple Choice Multi-select (MCMS)

- a. Answer key text is a comma separated list of numeric MC keys
- b. Should be limited to 2 keys per MCMS item
- c. An answer key of 1,4 would correspond to option A and D being correct in a 4 choice MC item

Sum Keyset

Number Of

Keyset Members

Integer Key

Value Key Weight

Integer Key

Value Key Weight

2. Inline Choice Multi-select

- a. Answer key text is a comma separated list of numeric keys
- b. Each numeric value is the answer key for the corresponding choice list
- c. Should be limited to 2 or 3 keys per item (one per score point)
- d. An answer key of 3,1,3 would correspond to keys of 3 in the first choice list, 1 in the second choice list and 3 in the third choice list

Sum Keyset	
Number Of	<input type="text" value="3"/>
Keyset Members	
Integer Key	
Value	<input type="text" value="3"/>
Key Weight	<input type="text" value="1"/>
Integer Key	
Value	<input type="text" value="1"/>
Key Weight	<input type="text" value="1"/>
Integer Key	
Value	<input type="text" value="3"/>
Key Weight	<input type="text" value="1"/>

3. Match Multi-select
 - a. Answer keys are a comma separated list of directed pairs of sources and targets
 - b. Source and target pairs are listed as source #/target # (e.g., 4 / 1 = source 4 goes in target 1)
 - c. No limits are specified for the number of source/target pairings in an item
 - d. A key set of 4 / 1,3 / 2 would correspond to source 4/target 1 and source 3/target 2

Sum Match Keyset	
Number Of	<input type="text" value="2"/>
Keyset Members	
Source Key	
Correct Target Id	<input type="text" value="1"/>
Source Key Id	<input type="text" value="4"/>
Source Weight	<input type="text" value="1"/>
Source Key	
Correct Target Id	<input type="text" value="2"/>
Source Key Id	<input type="text" value="3"/>
Source Weight	<input type="text" value="1"/>

4. Zones multi-select
 - a. Answer key text is a comma separated numeric list
 - b. Each key represents a zone key
 - c. No limits are specified to the number of selectable zones
 - d. A key of 1,3,4 corresponds to zones 1, 3, and 4 being keys

Point To Score Maps	
Point To Score Map	
Map Point Value	3
Map Item Score	2
Point To Score Map	
Map Point Value	2
Map Item Score	1
Sum Keyset	
Number Of	3
Keyset Members	
Integer Key	
Value	1
Key Weight	1
Integer Key	
Value	3
Key Weight	1
Integer Key	
Value	4
Key Weight	1

APPENDIX B: ELPA21 THEORY OF ACTION



Theory of Action

August 2014

"The contents of this document were developed under a grant from the U.S. 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."

The *English Language Proficiency Assessment for the 21st Century* (ELPA21⁹) is a consortium comprised of 11 states — Arkansas, Florida, Iowa, Kansas, Louisiana, Nebraska, Ohio, Oregon, South Carolina, Washington, and West Virginia — developing an assessment system designed to measure the performance of English language learners (ELLs) as they progress through their K-12 education and achieve college and career readiness.

Through the use of a screener and summative assessment, ELPA21 will support ELLs by determining initial proficiency and placement; identifying the need for reclassification or continued placement; providing information that can help guide instruction, nurture student growth, determine reclassification/exit status; and documenting accountability for the overall system and member states.

ELPA21 is unique in that it is designed to assess new English Language Proficiency Standards¹⁰ describing the how language is used by the rigorous content demands in each grade. As students practice language, they simultaneously interact with grade-level academic content. Increasing the expectations for the academic content that students must master in high school requires a parallel increase in expectations for English language acquisition. The ELP Standards describe these higher expectations by integrating language development with appropriate mathematics, language arts, and science subject matter. As ELs learn the academic uses of the English language, they are also exposed to the content knowledge necessary to be on track for college and career readiness.

Purpose of This Document

A Theory of Action (ToA) describes how ELPA21 will facilitate the move from current English language (EL) expectations and instructional practice to those necessary for ELLs to gain proficiency in the academic language used within math, science and ELA, and to ultimately become college- and career-ready. The ToA describes how our mission will be carried out, aligning intended assumptions with the organizational context of ELPA21. It connects strategy to action and identifies the multiple dependencies required for the successful implementation of our vision.

⁹ See ELPA21.org for additional information.

¹⁰ 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, CCSSO, 2013. Available at [http://www.ccsso.org/Documents/Final%204_30%20ELPA21%20Standards\(1\).pdf](http://www.ccsso.org/Documents/Final%204_30%20ELPA21%20Standards(1).pdf).

Mission: Acknowledging the diverse and rich language experiences ELLs bring to school, we recognize their English language proficiency is constantly growing. ELPA21 measures that growth based on the new ELP Standards and provides valuable information that informs instruction and facilitates academic English proficiency so that all ELLs leave high school prepared for college and career success.

Vision: To provide assessments that best measure English language learners' mastery of the communication demands of states' rigorous academic standards.

The ToA is grounded in research and evidence-based practice and describes what the consortium intends to achieve. It also connects strategy and actions to objectives and desired outcomes to fulfill the mission and vision of the organization and to address the following questions:

- How do we get from the current state to where we want to be?
- What steps need to be taken to improve EL teaching and learning?
- What evidence exists that the identified steps and strategies will be successful in the context of ELPA21?
- As a result of our actions, what can we expect to happen?
- How will these actions impact students, educators and schools?
- What results do we expect to see, and how will we measure them?
- What are indicators of success?

Theory of Action

The ELPA21 Theory of Action is based on a set of *core beliefs* and *foundational assumptions*. These distinguish ELPA21 from other ELP assessments. The *assessment system* reflects the synthesis and application of these core beliefs and foundational assumptions to specific goals that address emerging ELL needs and challenges and will result in the intended impact. *Planned actions* are the many complex steps and tasks that once complete, are expected to contribute to the impact of ELPA21. To evaluate the extent to which assessment objectives are met, *criteria for success* describe milestones and metrics that provide evidence of success throughout development and identify areas for additional refinement. Figure 1 describes the elements of the theory of action.

Figure 1. Elements of the Theory of Action.

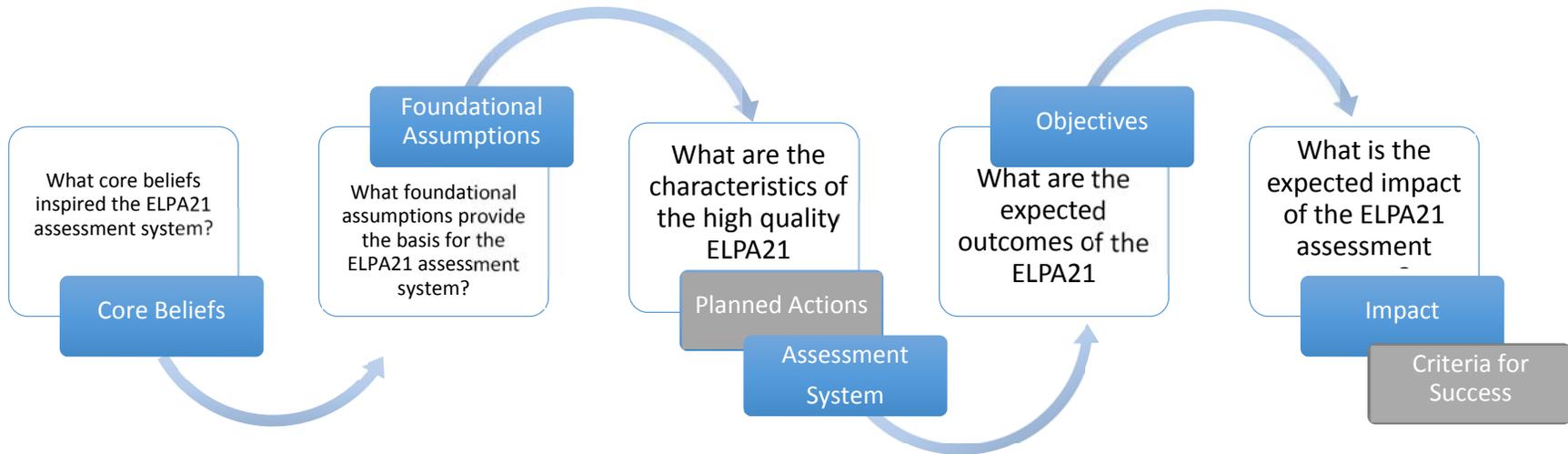
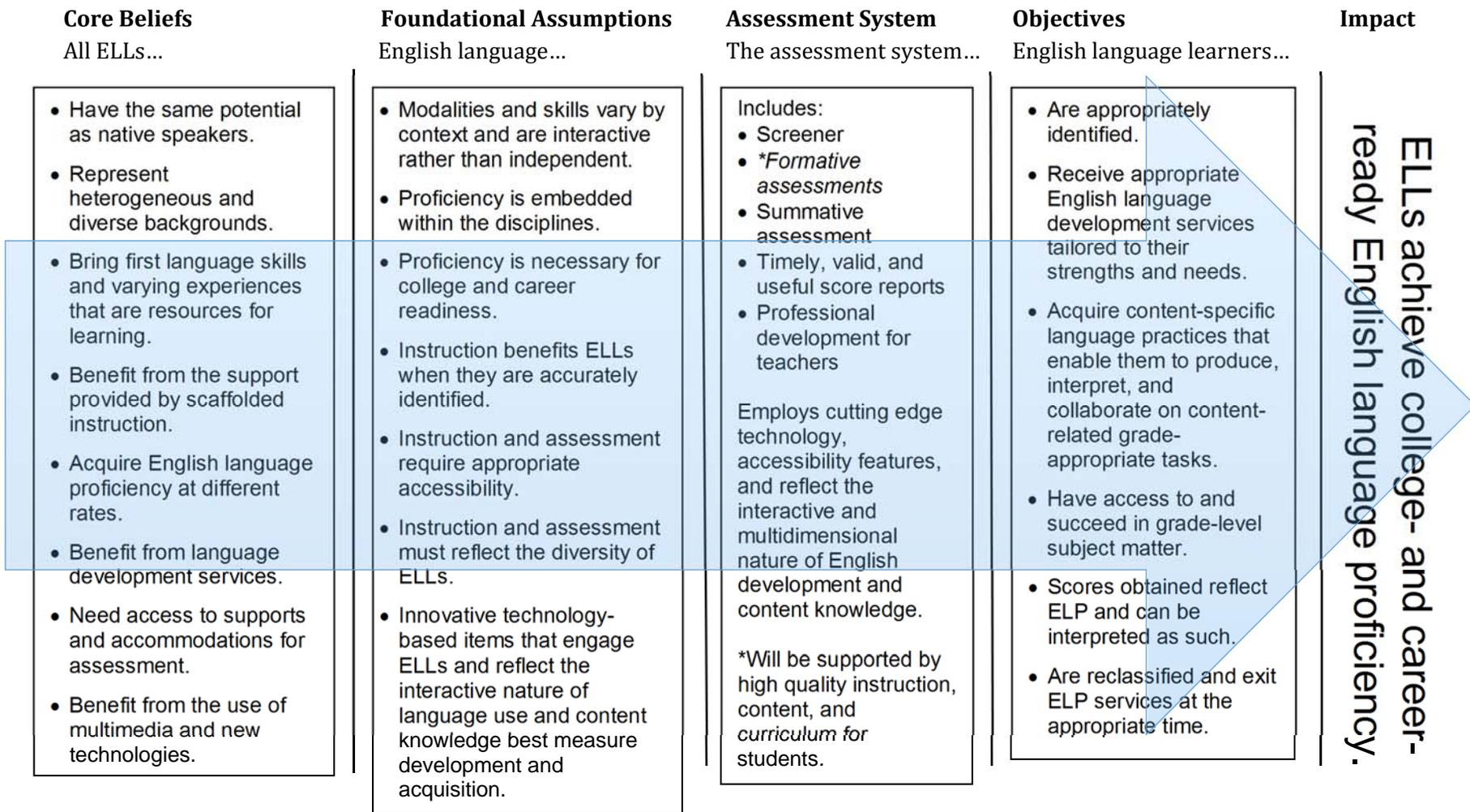


Figure 2 identifies each of these elements for the ELPA21 assessment system, and text following Figure 2 describes each.

Figure 2. ELPA21 Theory of Action.



Core Beliefs

At the heart of ELPA21 is a set of core beliefs. These beliefs drive the mission and vision, guide design and development, prioritize tasks and resources, and establish a new way of thinking about English language learning, instruction and assessment. Core beliefs also reflect the guiding principles of the ELP standards¹¹:

1. ELLs are a heterogeneous group, with physical, social, emotional, and/or cognitive differences, representing diverse social, educational, and cultural backgrounds. While they learn language at varying rates, all ELLs have the same potential as non-ELLs, and their diverse backgrounds are valuable resources for learning.
2. All ELLs are capable of making and demonstrating progress toward English language proficiency, and benefit from scaffolded instruction and language development services.
3. ELLs must acquire discipline-specific language practices that enable them to produce, interpret, and effectively collaborate on content-related grade-appropriate tasks. ELLs benefit from new technology and with the appropriate supports and accommodations, can make and demonstrate continual progress in the use of language.

Foundational Assumptions

The assessment design is guided by the three foundational assumptions concerning the nature of the English language. First, English language modalities (receptive, productive, and interactive) and domains (reading, writing, speaking, and listening) vary by context, and are interactive, rather than independent. Skills in each domain are developed interactively with, rather than in isolation from the other domains. The four domains are undeniably related to each other, and the standards and assessments reflect this.

Second, English language proficiency is not attained independently of the specific language processes that are embedded within each discipline. ELPA21 instruction and assessments are designed to align to the new English Language Proficiency Standards that correspond to the Common Core State Standards (CCSS) in ELA and mathematics¹² and the Next Generation Science Standards¹³. As a result, ELPA21 facilitates acquisition of the communication skills necessary for mastery of content standards.

¹¹ 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, CCSSO, 201, page 1-2.

¹² National Governor’s Association Center for Best Practices & Council of Chief State School Officers (2010). Common core state standards. Retrieved October 10, 2013 from <http://www.corestandards.org>.

¹³ ACHIEVE (2014). Next Generation Science Standards. Retrieved May 2, 2014 from <http://www.nextgenscience.org/next-generation-science-standards>.

Finally, English language proficiency, as described by the ELP Standards, corresponds to rigorous college- and career-ready standards. Accurate identification of ELL status is critical for students to receive the support necessary to become ready for postsecondary pursuits. Effective EL instruction and measures of progress towards mastery of the language must be accessible to, and reflective of, the diverse ELL population. Once proficient, ELLs are able to leave high school as prepared for college and career as their non-ELL peers.

Assessment System

Like the standards, the assessments focus on the *critical discipline-specific language skills* necessary for ELLs to become successful in school. This approach reframes language proficiency from “*What language does the student have?*” to “*What is the student able to do with language in the content areas?*” Comparing what students can do with what they need to be able to do provides more actionable and instructionally relevant information than does identifying a discrete set of knowledge and skills at a given point in time.

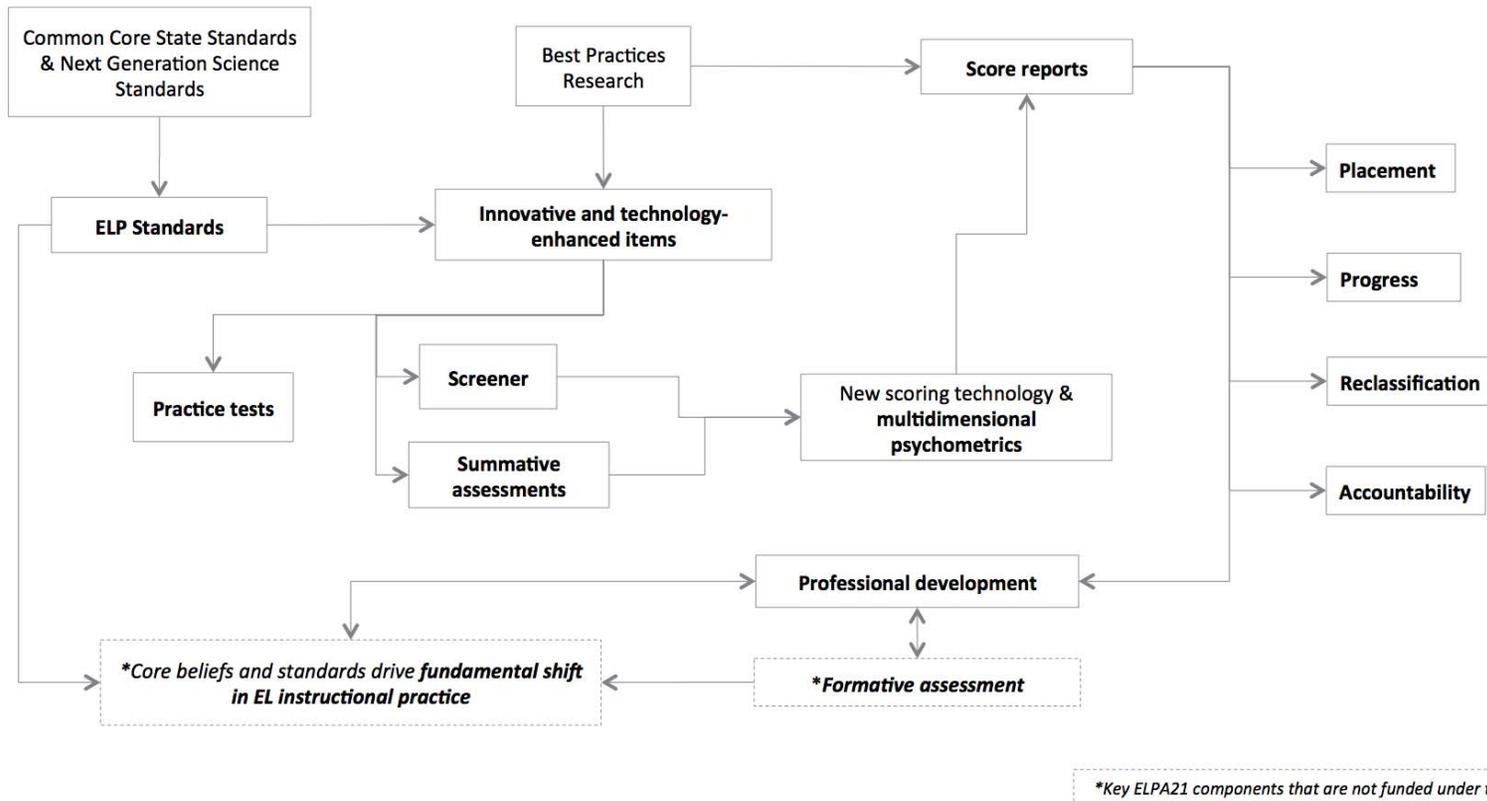
The assessment supports grade-level instruction and EL development that reflects the same change in thinking as the standards because ELLs can and should be supported in ways that allow them to become college- and career-ready at the same level as non-ELLs.

A new type of EL screener identifies potential ELLs by the ways they can use the language and not by the words and conventions they know. These students will benefit from a standards-based curriculum and formative assessment system that are not funded under the current assessment grant. This will include learning progressions and the interpretation of the current status of students based on the new expectations. Grade-band summative tests measure progress towards mastery of communication skills necessary for learning grade-level appropriate academic subjects. Measuring language acquisition regularly and with precision provides teachers and policy makers with information to make better instructional decisions for ELLs. Teachers can and should receive the support and student data necessary to provide appropriate and effective education to instruct students to more rigorous college and career readiness expectations.

Assessment results will reflect English language proficiency and can be interpreted as such. ELPA21 score reports are timely, and provide educators with useful results that inform individualized instruction.

Figure 3 describes the fully integrated ELPA21 system, including elements that are not a part of the assessment grant, such as formative assessment and standards based curriculum development and implementation.

Figure 3. ELPA21 Assessment System



Appendix B provides a more detailed diagram of the ELPA21 assessment system

Objectives

The purpose of ELPA21 assessments is described by four main objectives:

1. **PLACEMENT:** To determine the identification, current proficiency level, and appropriate placement of potential ELLs relative to grade appropriate performance standards.
2. **PROGRESS:** To monitor progress towards English proficiency for ELLs, describing individual and group strengths by domain and over time. Progress monitoring should meet multiple needs such as student placement and program exit, determining instructional needs of students and support needs of teachers, evaluating program effectiveness for subgroups of students, and adjusting educational programming and resources as needed.
3. **RECLASSIFICATION:** To determine proficiency relative to grade appropriate performance standards for reclassification purposes. Once proficient, students will have acquired the content-specific language practices that enable them to produce, interpret, collaborate on, and succeed in content-related grade-appropriate tasks.
4. **ACCOUNTABILITY:** To determine which districts are meeting accountability targets and identify schools in need of assistance.

In pursuit of these objectives, ELPA21 draws upon emerging technologies and innovative psychometric methods necessary to measure progress towards and mastery of the communication demands of rigorous academic standards.

Impact

The ultimate goal of ELPA21 is to remove language as a barrier to college and career readiness for ELLs. ELLs have the same potential as non-ELLs and must have the same expectation to leave high school proficient in the language necessary for college and career.

Planned Activities

The theory of action requires assessment design and development activities. Table A1 in the Appendix identifies the specific tasks necessary to complete the assessment system. Activities are designed and led by experts in computer-adaptive and fixed-form assessment, psychometrics, accessibility, item and task development, English language acquisition and development, standard setting, score reporting, and data use.

Criteria for Success

Numerous metrics throughout assessment development, implementation, and sustainability will determine the extent to which ELPA21 assessments meet stated objectives. The *ELPA21 Validity Plan* describes comprehensive plans to establish and document the reliability and validity of ELPA21 assessments throughout development, implementation, and sustainability. The ELPA21 Technical Report will describe the technical quality and rigor inherent in assessment design and development. The ELPA21 Sustainability Plan will describe the framework of organizational domains within the context of ELPA21. These domains will help define the parameters necessary to build the capacity for maintaining and enhancing the system leading to longevity and success. Table A1 in the Appendix describes criteria that will measure and establish project success.

Conclusion

ELPA21 has a unique opportunity to improve the way ELLs are prepared for entrance into college and careers. The ELP standards describe language proficiency as interactive in nature and embedded in grade-appropriate rigorous content knowledge. The ELPA21 instructional supports will guide educators teaching to these new standards, and ELPA21 formative assessments will reflect and inform enhanced instruction. The new ELPA21 assessments incorporate recent technological advances to measure, with precision, how students use language within academic contexts. All ELPA21 elements, when implemented, will make ELP more rigorous, more closely related to 21st century skills and rigorous content knowledge, and will redefine English language proficiency expectations, instruction, measurement, and outcomes.

Appendix A

Table A1 describes major tasks necessary for the development of an assessment system that will meet the four stated objectives. The *planned activities* for each objective are expected to result in the *anticipated outcomes*. The extent to which the anticipated outcomes are met can be determined by the *criteria for success*.

Table A1.Planned Activities, Outcomes, and Criteria for Success by Objective

Objective #1:	PLACEMENT: To determine the identification, current proficiency level, and appropriate placement of potential ELLs relative to grade appropriate performance standards
Planned activities:	<ul style="list-style-type: none"> *Compile evidence base and bring together national experts to identify and implement best practices in language learning, instruction, measurement, policy, accessibility, and assessment *Define grade-appropriate performance *Integrate standards into current classroom instructional practice *Applying Evidence Centered Design (ECD), develop innovative and technology enhanced test items that assess the interactive and content dependent nature of the standards *Identify and understand current diverse populations of actual and potential ELLs across all member states *Design items and delivery system to be accessible to increasingly diverse ELLs *Design, develop, and field test screener *Provide a valid and reliable screening measure that differentiates ELLs from non-ELLs *Implement scoring processes *Document data processing and psychometrics processes *Develop and deliver screener reports
Anticipated outcome:	<ul style="list-style-type: none"> *Potential ELLs participate in an efficient and effective screening process that is accessible to all students. *Results differentiate between ELLs and non-ELLs by measuring proficiency relative to grade-level performance standards. *Screener results determine program eligibility and identify instructional needs.
Criteria for success	<ul style="list-style-type: none"> *Evidence that classroom practice reflects new ELL expectations and practices *Evidence of balance between screener administration burden, technical quality, and useful information *Evidence of consistent administration, scoring, and classification processes *Evidence of classification accuracy *Evidence of correspondence between classroom observation and evidence and screener classification *Evidence of performance level validity, as established through standard setting *Evidence of scoring reliability *Evidence of accessibility to all students

Table A1.Planned Activities, Outcomes, and Criteria for Success by Objective, Continued

Objective #2:	Progress: To measure progress towards English proficiency for ELLs, describing individual and group strengths by domain and over time. Progress monitoring should meet multiple needs such as student placement and program exit, determining instructional needs of students and support needs of teachers, evaluating program effectiveness for subgroups of students, and adjusting educational programming and resources as needed.
Planned activities:	<ul style="list-style-type: none"> *Integrate emerging technology, best practice, and ECD *Establish technology specifications that are compatible with other assessments administered in member states *Design, develop, and validate summative assessment *Develop user support and guidance materials to ensure implementation with fidelity across member states *Report ELPA21 scores in ways that are useful and easily interpreted by intended audiences *Detect and report domain-level strengths and weaknesses to inform classroom instruction *Detect and report individual and aggregate proficiency *Develop and provide ongoing professional development to support educators in planning, implementation and improving standards-based curriculum and instructional plans *Measure and report growth in proficiency attainment over time
Anticipated outcome:	<ul style="list-style-type: none"> *A secure, logistically feasible platform that is interoperable, technically sophisticated, and that reliably delivers a summative assessment that is consistently implemented across member states. *It is accessible to all students and measures progress and proficiency with accuracy and precision. *Reports help teachers facilitate ELP in individual and groups of ELLs and help schools, districts and SEAs support teachers.
Criteria for success	Evidence listed for assessment objective #1, plus: <ul style="list-style-type: none"> *Evidence of implementation consistency within and across state *Score consistency with teacher observations of domain-level strengths and weaknesses *Evidence of validity and reliability *Evidence of scoring accuracy and reliability *Low-misclassification error *Minimal to no gaps in ELP attainment between subgroups *Plans for reliability and sustainability over time *Teachers indicate that assessment results are instructionally relevant and useful *Administrators affirm assessment results support resource allocation decisions and maximize accountability

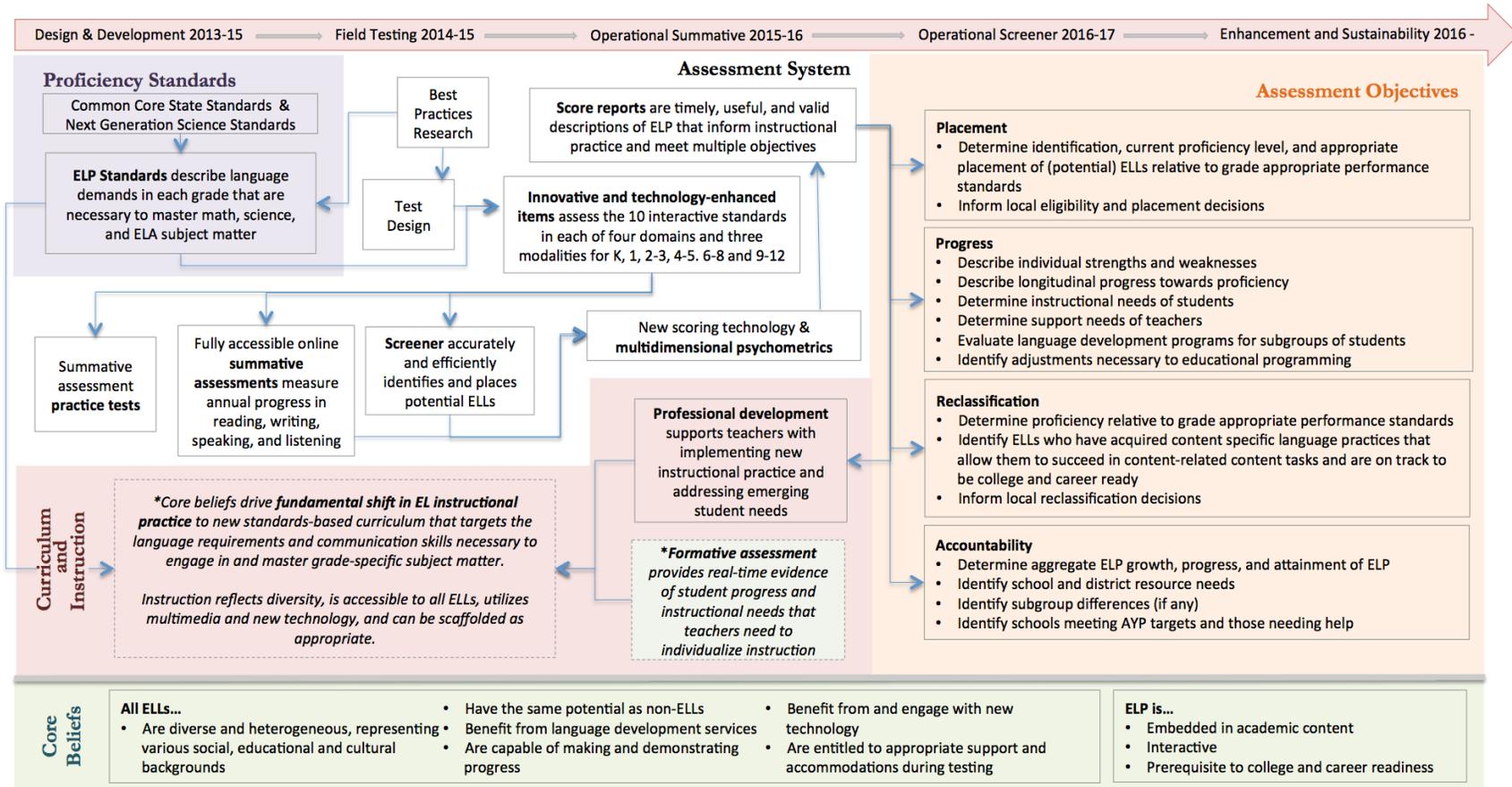
Table A1.Planned Activities, Outcomes, and Criteria for Success by Objective, Continued

Objective #3:	RECLASSIFICATION: To determine proficiency relative to grade appropriate performance standards for reclassification purposes. Once proficient, students will have acquired the content-specific language practices that enable them to produce, interpret, collaborate with others, and succeed in content-related grade-appropriate tasks.
Planned activities:	*Correctly reclassify ELL students who have become proficient in English at the level necessary to perform at grade-level
Anticipated outcome:	*Improved educator effectiveness and student achievement
Criteria for success:	*Evidence of score precision and reliability. *Low-misclassification error *Evidence of equivalence between screener and summative “proficient” scores

Objective 4:	ACCOUNTABILITY: To determine which schools are meeting accountability targets and identify school in need of assistance.
Planned activities:	*Aggregate scores at school, district, and state levels *Identify schools, ELL subgroups, or instructional areas needing additional resources or support *Provide policy-makers with the information necessary for high-level, high-stakes decisions
Anticipated outcome:	*Provide and model best practices for emerging bilinguals *Improved instruction for ELLs *Reallocated resources that address areas in need
Criteria for success:	*Increased percentage of schools meeting Annual Measurable Achievement Objectives (AMAOs)

Appendix B

Figure B1. Detailed ELPA21 Assessment System Diagram



APPENDIX C: PARTIAL CREDIT SCORING RULES VALIDATION REPORT



ELPA21 Partial Credit Scoring Rules Validation Report

Prepared by Emilie Pooler, Joyce Wang, and Bryan Doyle

Educational Testing Service

September 9, 2015



Introduction

This document summarizes and reports on the activities undertaken by Educational Testing Service (ETS) in 2015 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, and in collaboration with Questar Assessments Incorporated (QAI), to validate partial credit scoring rules¹⁴ for a corpus of test items within the ELPA21 assessment system's item pool.

The ELPA21 Consortium is a group of states organized to produce an assessment system to measure the language development of English language learners (ELLs). The ELPA21 assessment system is intended to be used by educational authorities in the consortium states for:

- determining initial identification of ELLs (via a 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.

In 2014, ETS, in collaboration with CCSSO and the ELPA21 Consortium, designed and developed a pool of test items based on the English Language Proficiency (ELP) Standards to form the basis of the ELPA21 assessment system. The ELPA21 item pool contains a number of innovative task types, including several technology-enabled (TE) item types. As part of the item design and development effort, ETS developed preliminary scoring rules for these item types, including scoring rules for awarding partial credit where appropriate (Hauck et al., in press, pp. 25–26). These scoring rules were conceived of as preliminary because ELPA21 and ETS recognized the importance of revisiting the scoring rules once field test data were available.

The purpose of the ELPA21 scoring rules validation effort was to analyze the psychometric results of the items on the field test that were eligible for partial credit scoring in order to validate the keys and the partial credit scoring rules, confirm preliminary rules were appropriate, and propose modifying or adding rules (i.e., determining which additional responses should be awarded full or partial credit) as appropriate.

¹⁴ The scoring rules referred to in this report are rules for the automated scoring of test items. In the field of assessment, such scoring rules are sometimes also referred to as rubrics. In this report, we have chosen to use the term scoring rules to reflect the important differences between this type of scoring activity and the rubrics used to guide human judgments made on individual student-constructed responses.

Process

Defining Partial Credit

As mentioned above, the ELPA21 item pool contains a number of innovative item types, including several TE item types. For ELPA21, TE items have been defined as “those computer-delivered items that include specialized interactions in the student response format or in the use of response data” (Hauck et al., in press, p. 25). Responding to a TE item requires a specialized interaction that can be more complex than responding to selected-response or text-entry (keyboarding/typing) items.

Because TE items can include more complex interactions, responses may provide more or different information about a student’s knowledge, skills, or abilities than a typical single-selection multiple-choice item. Whereas a typical single-selection multiple-choice item can only be scored dichotomously (correct or incorrect), some TE items can potentially be scored along a continuum of partially correct to fully correct depending on the number of interactions in an item. In addition to TE items, several multiple-select multiple-choice items for the ELPA21 assessment were also designed with the intent to apply partial credit scoring. For the purposes of ELPA21, partial credit scoring is defined as a proportion of credit awarded for a student response to a TE or multiple-select multiple-choice item that demonstrates some, but not total, success in completing the task.

Criteria for Partial Credit Scoring Rules

The partial credit scoring rules for ELPA21 were not based on item complexity (e.g., the number of interactions required to respond to an item) or item difficulty. Instead, the partial credit scoring rules were related to the test construct and were based on distinctions among the ELPA21 proficiency-level descriptors (PLDs). That is, responses awarded a proportion of credit would provide evidence of students’ level of English language proficiency, as described in the PLDs.

Additionally, the scoring rules validation required that responses to proposed partial credit items have the following characteristics.

- The score obtained for each item should be positively related to student proficiency; that is, the student’s overall score should be positively correlated to the TE item score.
- The score levels described in the scoring rules should be supported by actual student responses; that is, the distinction between score levels should be clearly demonstrated by student responses, and each of the distinct score levels must be observed in the empirical data.

In addition to validating the draft scoring rules, it was also important to examine the empirical data to refine or adjust those draft scoring rules. When the scoring rules were first defined, they were based on ETS content experts’ assumptions about likely multiple levels of proficiency, matching the PLDs defined for ELPA21 (see below). In order to validate the original assumptions of the scoring rules and support further refinement of these rules, students’ actual response patterns needed to be examined.

Developing Preliminary Partial Credit Scoring Rules

As part of the initial development of the item pool, ETS content experts reviewed each item in the pool to determine which items might be eligible for partial credit by applying the criteria described above (i.e., distinctions between score points should be construct-related). Since the ELPA21 construct is based on a set of standards and associated PLDs¹⁵, the ELP Standards and PLDs served as references in making determinations about whether distinctions in performance would be related to the construct.

ETS content experts made judgments as to whether the student responses to an item could address two or more PLDs of the aligned standard(s) and whether the responses would, therefore, provide distinct information regarding the student’s level of English language proficiency. In cases where different aspects of the construct could be revealed by a student’s partial response, scoring rules for awarding partial credit were created. The rules were created on an item-by-item basis. Unlike a set of rules that would be based solely on an item type or number of interactions, scoring rules based on the construct assessed by each item must be made on a case-by-case rather than a global basis.

Once scoring rules for awarding partial credit were drafted, the items and associated rules were reviewed by the ELPA21 Item Acquisition and Development Task Management Team (IAD TMT). The IAD TMT reviewed the items and accompanying rules and provided input. A final list of items eligible for partial credit scoring was created.

The Corpus of Partial Credit Items

Of the 1,138 TE items in the ELPA21 pool, 89 were determined to be eligible for partial credit scoring following the criteria outlined above. The item types included 43 “drag-and-drop¹⁶” items and 46 “drop down¹⁷” items. An additional 17 multiple-select multiple-choice items were included in the corpus for an initial total of 106 items to validate. Thirty-six of the drop down items, which contained three interactions each, were set aside for separate review, leaving the corpus to be validated at 70.

Before the field test, scoring rules for the corpus of 70 items, reflecting the IAD TMT’s input, were entered into the ETS Item Banking and Information System (IBIS) database for each eligible item. These initial scoring rules were considered preliminary and were to be evaluated and validated with actual student data once the field test had been administered.

¹⁵ The ten ELP Standards are organized according to how language skills correspond to English language arts and literacy, mathematics, and science standards in six grade bands: Kindergarten, Grade 1, Grades 2–3, Grades 4–5, Grades 6–8, and Grades 9–12. In addition, each of the ELP Standards includes PLDs at five proficiency levels. These PLDs describe target performance for ELLs 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 band. That is, each standard is further defined with grade-appropriate expectations at each of the five proficiency levels.

¹⁶ In drag-and-drop items, students respond by dragging and dropping choices (“sources”) into the appropriate locations (“targets” or “drop zones”).

¹⁷ In drop down items, the item contains one or more blanks, and the student must fill in each blank with a choice from its list.

Field Test Administration

After the pool of field-test-ready ELPA21 items was developed, ETS handed the pool off to QAI, the field test delivery vendor. QAI then prepared the items for field test administration, administered the ELPA21 field test (in February–March 2015), and scored the field test. Initial scoring of the partial credit items was done by QAI following the preliminary scoring rules developed by ETS.

Field Test Data for Partial Credit Items

Response data for the items eligible for partial credit scoring, provided by QAI, were crucial to the scoring rules validation effort.

Classical test statistics were used in the validation effort. Item Response Theory (IRT) scaling and the associated IRT statistics were not available at the time of the validation effort because of the smaller than expected scale of the field test and the timing of the scoring rules validation. For Kindergarten and Grade 1, items were administered in test forms by domain (i.e., reading, writing, listening, and speaking). For grades 2–12, the field test design called for items to be administered in item “blocks;” each block was designed to take half the time allowed for an entire domain form so provided fewer score points than a full form would. Item form or item block scores for each student was used as a proxy for student proficiency.

The statistics provided by QAI for the scoring rules validation effort consisted of item-level P-values, item-item block score correlations, average item block scores for each score level, and average item block scores for each observed response pattern. In addition to item statistics, item metadata such as item ID, item set ID, maximum possible score, and intended key for full score response were also included in the data. Appendix B shows the data layout for the item response data file. QAI provided a separate data file for item domain/block summary statistics. The domain/block summary data contained information about number of items for each domain for Kindergarten and Grade 1, number of items for each block for grades 2–12, minimum/maximum domain/block scores, and mean and standard deviation of domain/block scores.

Unlike standard single-selection multiple-choice items, each partial credit item required students to enter multiple responses to completely respond to the prompt. As a result, QAI captured the students’ responses as strings rather than a single response. A typical four-option multiple choice has five possible responses, A, B, C, D, or omit. ELPA21 TE items, however, have up to ten possible responses for each response part. This results in items with up to 32,000 permutations of possible student responses. For the ELPA21 field test, the number of observed response permutations ranged from 7 to 339 with an average of 23. These large numbers of responses made item analysis a complex endeavor.

Effective Key

ETS established an “effective key” from the original observed response data for multiple-interaction items eligible for partial credit scoring. Depending on the task type, the order of the keys in the response



string was either fixed (e.g., for a fixed key of A,B,C a student response string of B,A,C is incorrect) or order independent (i.e., the order in which a student provides responses does not matter). Because students could enter multiple values for each response part, ETS analyzed each pattern of student responses to identify equivalent responses (e.g., for a fixed key of A,B,C responses A,D,C and A,A,C would be equivalent responses). That is, there can be more than one way for students to come to a response pattern that is effectively the same. Table 1 demonstrates how student response strings are grouped into the same effective key. When the order of the responses matters for students responding to an item with three elements, the response strings of B,blank,D; B,A,D; B,C,D; and B,D,D indicate correct responses for the first and third element of the item; therefore the effective key for those responses are B,blank,D.

Table 1: Example of Effective Key

Item Scenario	Key	Student Response String	Effective Key
Where the order of the responses matters, the effective key identifies which elements of the key the student responded to correctly.	B,B,D	B,_,D	B,_,D
		B,A,D	B,_,D
		B,C,D	B,_,D
		B,D,D	B,_,D
Where the order of the responses does not matter, the effective key identifies which responses the student correctly selected.	B,B,D	B,D,C	B,_,D
		A,D,B	B,_,D
		B,C,D	B,_,D
		B,D,B	B,B,D

The effective key concept is used to aggregate and reformat student responses and allow ETS to present student response data in a more focused, easily comprehensible format. Table 2 below represents how the 64 rows of response data (shown in Appendix C) were aggregated to seven rows of summary data by effective key. Summary data were provided to a panel of experts and stakeholders established by the ELPA21 Consortium to validate the scoring rules.



Table 2: Example of Summary Data by Effective Key

		P-value	Average Item Score	Score Point	# of Students	Effective Key	# of Students with Effective Key	Average Domain Block Score
Item	VHXXXX XX	0.54	1.63	0	103			
Max Points	3			1	188	3,_,_	68	2.191324
Student Count	676					_,1,_	50	2.02
Full Answer Key	3,1,2					_,_,2	70	2.369571
				2	243	3,1,_	57	3.121053
						3,_,2	109	3.421376
						_,1,2	77	3.555195
				3	142	3,1,2	142	4.67

In addition to producing effective keys, ETS used the average domain score for all of the items to calculate the average domain score for the items with an effective key. ETS also carried out a response analysis that identified possible issues with how students performed on the items (e.g., for many drag-and-drop items, students disregarded directions and dragged a source into every available target rather than only dragging the number directed). This analysis enabled the panelists to focus on student response behaviors and guide decisions on how the items should be scored.

As an initial part of the analysis, ETS conducted a routine key-check review to ensure the score keys were correct. (If the score keys were inaccurate, it would negatively impact the conclusions made for the partial credit scoring rules.) After applying the corrected keys to the item response file, ETS was able to compute correct P-values, number of students for each effective key, and mean scores for each effective key. Items with multiple correct response patterns in the observed data were also flagged for review since it was expected that only a single correct response pattern should be allowed. A more in-depth discussion on items with multiple correct response patterns is provided in the Panel Meeting and Discussion section.

Ordinal relationship between score levels and the average mean block/form scores at the individual item level were observed for all items in the study, which indicated the preliminary scoring rules were functioning reasonably well.

Panel Meeting and Discussion

A key stage in the validation of the scoring rules was the presentation of the preliminary scoring rules, the field test results, and recommendations for confirming or adjusting the scoring rules to a panel of experts and stakeholders established by the ELPA21 Consortium.

The ELPA21 Consortium Council was given the opportunity to invite up to two representatives per state to participate, and preference was given to those with measurement and/or a background in ELL education. In the end, 14 people participated in the panel to validate the scoring rules, including at least one representative from each of the nine current ELPA21 states (Arkansas, Iowa, Kansas, Nebraska, Ohio, Oregon, South Carolina, Washington, and West Virginia). Along with those voting panelists, additional participants participated as observers or discussants including staff from ELPA21, ETS, and Questar. (See Appendix A for a list of participants in the panel meeting.)

The panel meeting was held via WebEx over the course of two consecutive days and was facilitated by the ETS assessment development and psychometric leads, along with input from the ELPA21 IAD TMT lead.

On the first day of the panel, the criteria and process for creating partial credit scoring rules were reviewed. Then the panel began the review of the 70 items eligible for partial credit scoring. For each item, the following information was presented or made available:

- a screen shot illustrating how the item was administered to students in the field test,
- the PLDs the item was intended to measure,
- the preliminary scoring rule for the item, and
- response data from the field test administration.

Screen shots. First, each item was presented as a screen shot showing how the item appeared during field testing. It was important that the panel be able to review response data and make their judgments with an understanding of how the item actually appeared to the test takers. These screen shots were in color and, when necessary, included a set leader or stimulus and the item itself.

Performance Level Descriptors. In order to convey the construct each item was intended to measure, PLDs for the standard(s) each item was aligned to were made available to the panel. Because the criteria for awarding partial credit was to be based on potential differences in achievement of the construct, the intended construct for each item needed to be made available to the panelists.

Chart 1 below provides an example of how the five levels of PLDs were presented for a given standard.

Chart 1: Example of Proficiency Level Descriptors

<p>ELP Standard 1 Listening Subclaim: 1L The English language learner constructs meaning from oral presentations of literary and informational text. The learner can:</p>				
Level 1	Level 2	Level 3	Level 4	Level 5
use a very limited set of strategies to: <ul style="list-style-type: none"> identify a few key words and phrases from read-alouds, simple written texts, and oral presentations 	use an emerging set of strategies to: <ul style="list-style-type: none"> identify some key words and phrases identify the main topic or message/lesson from read-alouds, simple written texts, and oral presentations 	use a developing set of strategies to: <ul style="list-style-type: none"> identify the main topic or message answer questions retell some key details from read-alouds, simple written texts, and oral presentations 	use an increasing range of strategies to: <ul style="list-style-type: none"> determine the main idea or message identify or answer questions about some key details that support the main idea/message retell a variety of stories from read-alouds, written texts, and oral presentations 	use a wide range of strategies to: <ul style="list-style-type: none"> determine the main idea or message tell how key details support the main idea retell a variety of stories from read-alouds, written texts, and oral communications

Scoring Rules. The preliminary scoring rule for each item was also presented to panelists. In some cases, the rule was fairly straightforward and was simply written out (e.g., “1 point per correct interaction”). In cases where the rule was more complex, it was presented in tabular form, as shown in Table 3.

Table 3: Example of Scoring Rule Presentation in Tabular Form

Number Correct	Points
0, 1	0
2	1
3	2

Response Data. Finally, a summary of the response data for each item was presented as shown in Table 2. In addition to the response data that was provided to the entire panel, additional information was available to the panel if further discussion was warranted (e.g., item-total correlations, average raw scores at each score point). Such information was not routinely provided to panelists, mainly because of the small number of students responding to the item or a small number of items in each block.

Discussion. For each item eligible for partial credit scoring, panelists were given time to consider the content of the item, the scoring rules, and the response data. The panelists could comment on the item content and data, and they could ask questions for the entire panel to consider. The panel discussed each item and its accompanying scoring rules and response data.

After each item was reviewed and discussed, the meeting facilitator called for a vote to approve the scoring rule for the item. In some cases, new scoring rules were proposed during discussion, and the new scoring rules were voted on. In other cases, the panel decided not to vote on a rule until the item response data could be reviewed in more detail.



In addition to feedback on the partial credit scoring rules, a number of other comments were made by panelists. Issues determined to be outside the scope of the scoring rules validation effort were recorded as “parking lot” issues. These issues were compiled in a document by ETS and shared with ELPA21 for further consideration.

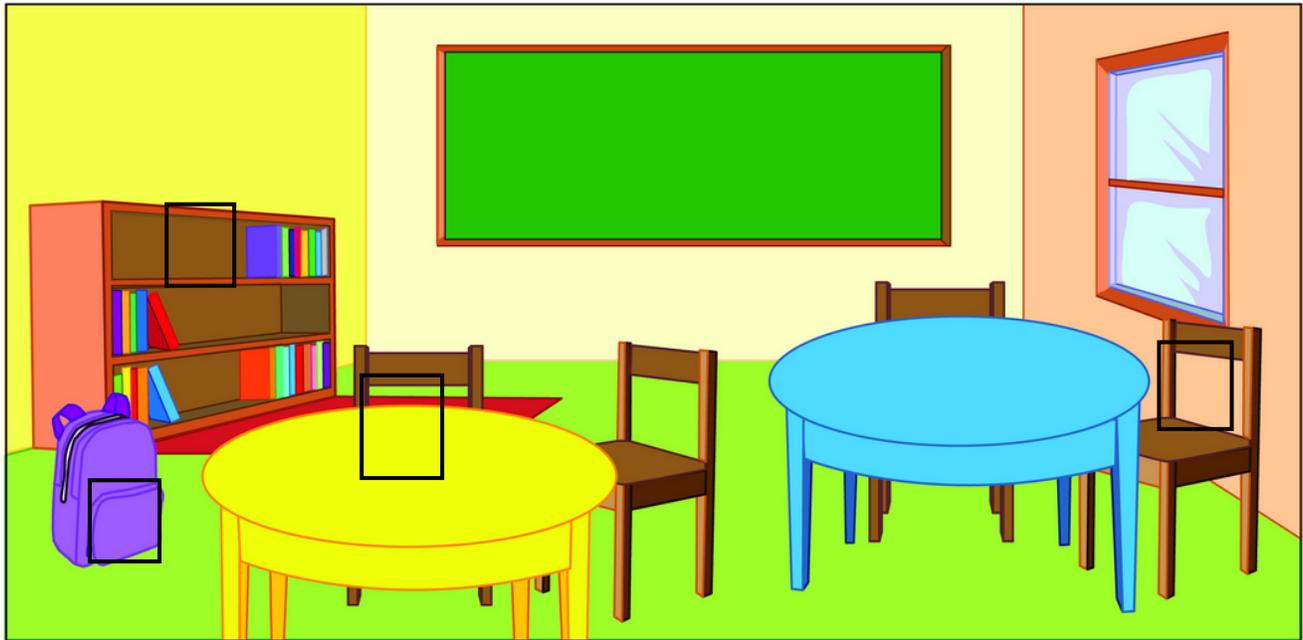
There were many rich discussions that took place during the scoring rules validation. Three topics of particular interest generated the most discussion during the WebEx:

- students’ providing more responses than called for in the instructions,
- precluding the possibility of automatic credit on items with multiple keys,
- partial credit scoring of order strings in sequence items.

Each of these topics is summarized in turn below.

Additional Responses. In 9 of the 43 drag-and-drop items, more objects and “drop zones” were provided than were needed to respond to the item correctly. That is, a student was asked to select and place an object or objects into a specific space, or drop zone, but there were additional drop zones in which to place the objects. These additional drop zones were intended to serve a purpose analogous to that of distractors, or incorrect answer choices, in a traditional multiple-choice item. For example, in this item type, a listening question might show a picture of a classroom, with the item prompt directing students to select an image of an object to place in a specific location in the classroom. A student hearing the prompt “Put the book on the shelf.” would first need to make a selection from a group of objects (e.g., book, ruler, pencil). Then the student would need to select a location to place the object. In addition to the shelf, a number of other spaces are available for placing the book, as indicated by the outlined drop zones in Figure 1 (e.g., table, chair, backpack).

Figure 1: Sample Item with Multiple Zones and Sources



In the 9 items where additional drop zones were provided, some students completed the intended task but then did more than the task called for. That is, the students would drag the book to the shelf, and then place the additional objects into the remaining drop zones. For example, the students might put the pencil on the table and the ruler on the backpack, even though students were not prompted to do anything with these two objects.

The first item in which response data revealed that students performed additional interactions not called for by the instructions was discussed at length. After much discussion, a number of different rules were proposed and considered.

Proposed Rules. Some proposed rules disregard the number of additional objects the student places on the picture. That is, once the student has satisfied the condition specified for the score point, any additional interactions are not considered in scoring. Other proposed rules would penalize students for going beyond what is called for in the instructions for the item.

Discussion centered on the construct being measured, the age of the test takers, and the presentation of the item on the delivery platform. The intended construct for these ELPA21 items under review is not



whether students can follow directions; it is whether they can identify key words and phrases in main messages in oral communication. In addition, this item was administered to young learners. It might be seen as intuitive for children in lower grades especially to place all objects into all open locations. Since the drop zones were visible to the test takers, they might understand their task as moving objects to all drop zones. In addition to discussion about the construct being measured, data were available for review. The block scores for those students who placed extra objects in drop zones was available to the panelists. In general, data showed that students who placed extra objects in additional drop zones performed at a similar level to the students who completed only the intended task.

In the end, based on both the construct considerations and data reviewed, the panel agreed that no penalty would be assessed to students who provide additional responses. A vote was taken and the decision was unanimous.

Precluding Automatic Credit. A second issue that generated discussion centered on the items in which test takers might receive partial credit by simply following the directions for an item. For example, in an item in which students are asked to choose three correct answers out of five options, if a student chooses any three options, one of the those three will be correct by default. In those cases, the panel favored rules that did not award credit for providing one correct response. Similarly, if an item asks students to choose four answers out of six options, a student would get two correct by just completing the task. In that case, the panel favored awarding partial credit only for students who answered three parts of the item correctly and full credit for students who answered all four correctly.

Scoring Order Strings. A third area of discussion centered around sequence items, which required students to put information into a correct sequence. The preliminary scoring rules allowed for responses to be scored as correct only if each selection was in the correct position of the sequence (1,2,3,4). Partial credit scoring could only take into account whether the correct sources were placed in the correct targets or drop zones and not whether a partial order string—that is, two or more selections in correct consecutive order or sequence—was placed in the incorrect target or drop zone. For example, if a student placed the sources into order 1,3,2,4, the student would be given partial credit for correctly sequencing the first and final part of the item. However, if a student placed the sources into order 2,3,4,1, zero points (no partial credit) would be awarded based on the preliminary scoring rules, even though there is a partial order string (2,3,4) in the correct sequence. There were six sequence items in the corpus, and the panel voted to put the scoring rules for these items on hold in order to investigate further. The panel suggested that a number of responses patterns be investigated to determine if and how certain responses patterns might be related to the test construct.

Results

Panel Results

In the end, the panel voted on the partial credit scoring rules for 70 items. For 60 items, the scoring rules were approved as originally drafted. For four items, a new scoring rule was created and approved. Six items have been put on hold for further review before a rule is confirmed.

**Generalizable Rules**

Two conceptual issues arose from the panel discussion that are now reflected as part of the modified scoring rules.

- Students who provide additional responses for all items reviewed by the panel are not penalized for doing so. This rule should be applied on an item-by-item basis.
- If it is possible to get one (or two) response(s) correct simply by completing the task, the scoring rule should not result in automatic credit. However, this rule might need to be reviewed on an item-to-item basis in the future so that the test construct can be considered.

A third area that is likely to result in a conceptual rule has to do with the six sequence items that were put on hold. The scoring rules will be confirmed once additional response patterns and associated data are analyzed.

Directions for Future Research

The ELPA21 partial credit scoring rules validation effort left open a number of possibilities for future research and for future item development.

- Investigate construct-based rules for partial credit scoring. It is worth exploring how these complex TE and multiple-selection multiple-choice items are scored in other programs.
- Investigate how to take into account any additional responses a student might provide beyond what is required by the task. How should these be scored? Can revised test directions reduce the chances of this occurring? Can TE items be constructed differently to reduce the chances of this happening?
- Consider what methods of data analysis are best suited for complex TE items. Exploring best practices for analysis of responses with so many possible permutations is timely in the era of innovative, computer-based testing.
- Continue to research appropriate measurement models. Panelists were tasked with deciding on the final scoring rules by taking into account the alignment of item responses to PLDs and the empirical field test data associated with an item. The panelists did not consider whether specific measurement models were more appropriate for certain item types. A few of the drop down items determined to have four score levels (i.e., score of 0, 1, 2, and 3) could alternately be considered as three separate dichotomous items. For these items, the alignment to the PLDs can be interpreted more broadly in order to consider them as dichotomous. Fitting these items with a three-level polytomous IRT model versus three separate dichotomous items would be an interesting research topic to pursue.

References

Council of Chief State School Officers. (2014). *English Language Proficiency (ELP) 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.

Hauck, M.C., Pooler, E., and Anderson, D.P. (May 15, 2015). *ELPA21 Item Development Process Report*. Washington, DC: CCSSO.

Appendix A: List of Panel Participants

State Representatives (Voting Panelists)

- Mimi Brailsford (South Carolina State Department of Education)
- Zack Conrad (Kansas State Department of Education)
- Tom Deeter (Iowa Department of Education)
- Phyllis Farrar (Kansas State Department of Education)
- Susan Gray (Arkansas Department of Education)
- Kim Hayes (Washington Office of Superintendent of Public Instruction)
- Tom Hirsch (Psychometric Consultant to Washington Office of Superintendent of Public Instruction)
- Mami Itamochi (West Virginia Department of Education)
- Jobi Lawrence (Iowa Department of Education)
- Alan Lytle (Arkansas Department of Education)
- Abdinur Mohamud (Ohio Department of Education)
- Terri Schuster (Nebraska Department of Education)
- Steve Slater (Oregon Department of Education)
- Kurt Taube (Ohio Department of Education)

Non-Voting Participants

- Wes Bruce (ELPA21)
- Mark Hansen (ELPA21)
- Margaret Ho (CCSSO)
- Cat Still (ELPA21)
- Phoebe Winter (ELPA21)

Observers

- Jared Agrimson (Questar)
- Stephanie Phahlert (Questar)
- Amy Snyder (Questar)
- Lei Yu (Questar)



Appendix B: Data Layout from QAI

2015 ELPA21 Item Response File to ETS

Revised: 05/20/2015

Column	Field Length	Field Name	Column Name	Field Description	Acceptable Values
Item Record Type					
A	10	Item UIN	Item	This is the Item UIN.	Alphanumeric.
B	10	Set Leader UIN	SetLeader	This is the set leader item UIN that the Item UIN belongs to in the test. It is found in the spreadsheet (column C) that has max score points.	Alphanumeric.
C	2	Maximum Score Points	MaxPts	This is the maximum value of score points for this item.	Numeric
D	5	Count of Students Responding	StudentCnt	This is the count of students who responded to this record.	Numeric
E	10	P Value	Pvalue	P Value - to be filled in by ADP	Numeric (Blank until filled in by ADP)
F	10	Average Item Score	AvgItemScore	Average Item Score - to be filled in by ADP	Numeric (Blank until filled in by ADP)
G	10	Item Domain/Block Correlation	ItemDmn_BlkJCorr	Item Domain Correlation - to be filled in by ADP	Numeric (Blank until filled in by ADP)
H	2	Number of Possible Scores	PossScoreCnt	Item Domain Correlation for K and 1- to be filled in by ADP Item Block Correlation for 2-3 and Up - to be filled in by ADP [Different test design was used: fixed forms for K and 1 and block design was used for 2-3 and up.	Numeric



				In a block design, a domain form (e.g., that for the Reading domain) was made of any 2 block combinations.]	
Score Value Possibilities 1 to N					
I	2	Item Score Value	ItemScoreVal	This is the first score value for this record and represents all responses that received a zero score.	Numeric
J	5	Count of Students with this Score	StudentCntSco	This is the count of students who received this score for this item.	Numeric
K	3	Average Domain/Block Score	AvgDmn_Blksco	This is the average domain or block score for all students who received this score. To be filled in by ADP.	Numeric (leave blank for ADP)
L	3	Number of Pattern Permutations	PattPermCnt	This is the number of permutations of responses for this item receiving this item score. This does not represent all permutations, but only those permutation patterns that students responded to.	Numeric
Permutation Patterns for the first Score Value Permutations 1 to N					
M	10	Permutation Pattern	Response_String	This field is the student response in the form of a string. Since there will be more than one response to an item, there will be separators in the form of commas (",") and backslashes ("/").	Alphanumeric, including special characters of underscore ("_"), comma (","), and back-slash ("/") Examples: + 1,3 represents two responses that

					<p>are drop down like items.</p> <p>+ 1/3 represents a response for drag and drop like items</p> <p>+ 1/3, 3/4 represents two responses for drag and drop down like items</p> <p>+ NULL or "_" (by itself) represents no response</p> <p>+ 1,_ represents two possible responses, but only the student only responded to one of the parts of the item</p>
--	--	--	--	--	---

N	5	Count of Students with this Pattern	StudentCntPatt	This is the count of students who responded with this pattern for this score for this item.	Numeric
O	3	Average Domain/Block Score	AvgDmn_BlkscoPatt	This is the average domain or block score for all students who responded to this pattern. To be filled in by ADP.	Numeric (leave blank for ADP)



Appendix C: Example Response Data Used for Effective Key

Item	MaxPts	Student Cnt	Pvalue	ItemDmnBlkC orr	ItemScoreVal	AvgDmnBlkSco	PattPermCnt	ResponseString	FullAnsKey	StudentCntPatt
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	._._	3,1,2	45
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	._._1	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,2,1	3,1,2	4
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,2,4	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,3,1	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,3,3	3,1,2	4
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,3,4	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,4,1	3,1,2	10
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	1,4,4	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2._._	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,2,1	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,3,1	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,3,3	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,3,4	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,4,1	3,1,2	9
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,4,3	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	2,4,4	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	4,2,1	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	4,3,1	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	4,3,4	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	4,4,1	3,1,2	5
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	4,4,3	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	0	0.62	23	4,4,4	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,1._	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,1,1	3,1,2	12
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,1,3	3,1,2	4
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,1,4	3,1,2	3
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,2,2	3,1,2	9
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,3,2	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	1,4,2	3,1,2	18
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,1._	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,1,1	3,1,2	14
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,1,3	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,1,4	3,1,2	4
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,2,2	3,1,2	6
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,3,2	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	2,4,2	3,1,2	25
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3._._	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3._.1	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,2._	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,2,1	3,1,2	3
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,2,3	3,1,2	6
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,2,4	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,3,1	3,1,2	5
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,3,3	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,4,1	3,1,2	17
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,4,3	3,1,2	11
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	3,4,4	3,1,2	20



VHXXXXXX	3	676	0.54	0.84	1	2.21	31	4,1,1	3,1,2	4
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	4,1,3	3,1,2	3
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	4,1,4	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	4,2,2	3,1,2	2
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	4,3,2	3,1,2	1
VHXXXXXX	3	676	0.54	0.84	1	2.21	31	4,4,2	3,1,2	6
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	1,1,2	3,1,2	58
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	2,1,2	3,1,2	11
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	3,1,1	3,1,2	31
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	3,1,3	3,1,2	15
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	3,1,4	3,1,2	11
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	3,2,2	3,1,2	21
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	3,3,2	3,1,2	13
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	3,4,2	3,1,2	75
VHXXXXXX	3	676	0.54	0.84	2	3.40	9	4,1,2	3,1,2	8
VHXXXXXX	3	676	0.54	0.84	3	4.67	1	3,1,2	3,1,2	142

Appendix D: Sample Slides from Panel Meeting

Sample Grade 1 Listening Item
 Narrator: "Look at this classroom. Listen to the teacher. Follow the teacher's directions."
 Teacher: "Put the ruler on the table. Put the book on the backpack."

Proficiency Level Descriptors (PLDs)

ELP Standard 1				
Listening Subclaim: 11 The English language learner constructs meaning from oral presentations of literary and informational text. The learner can:				
Level 1	Level 2	Level 3	Level 4	Level 5
use a very limited set of strategies to: <ul style="list-style-type: none"> Identify a few key words and phrases from read-alouds, simple written texts, and oral presentations 	use an emerging set of strategies to: <ul style="list-style-type: none"> Identify some key words and phrases Identify the main topic or message/lesson from read-alouds, simple written texts, and oral presentations 	use a developing set of strategies to: <ul style="list-style-type: none"> Identify the main topic or message Answer questions Retell some key details from read-alouds, simple written texts, and oral presentations 	use an increasing range of strategies to: <ul style="list-style-type: none"> Determine the main idea or message Identify or answer questions about some key details that support the main idea/message Retell a variety of stories from read-alouds, written texts, and oral presentations 	use a wide range of strategies to: <ul style="list-style-type: none"> Determine the main idea or message Tell how key details support the main idea Retell a variety of stories from read-alouds, written texts, and oral communications

Confidential and Proprietary. Copyright © 2013 by Educational Testing Service.

Sample Response Data Table

		Pvalue	AvgItemScore	Score Point	# of students	Effective key	# of students with effective key	AvgDmnBkSco
Item	VHYYYY Y	0.71	1.42	0	44		44	
MaxPts	2			1	93	2	48	11.68625
StudentCnt	313					4	43	12.04244
FullAnsKey	2,4			2	176	2,4	70	14.157