Oregon Department of Education

2013–2014

ELPA

Oregon's Statewide Assessment System

Annual Report

Submitted by American Institutes for Research®
September 15, 2014

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

1. OVERVIEW

1.1 Purpose of ELPA

The purpose of Oregon's English Language Proficiency Assessment (ELPA) is to assess academic English ability in reading, writing, listening, speaking, and comprehension for English Language Learners (ELLs) enrolled in Oregon public schools in grades K–12.

As part of the Elementary and Secondary Education Act (ESEA) and the previous No Child Left Behind Act (NCLB) enacted in 2001, states must annually measure and report progress toward and attainment of English language proficiency by ELLs enrolled in public schools. Under ESEA, states must develop English Language Proficiency (ELP) content standards linked to content standards including those for English Language Arts (ELA). The Oregon English Language Proficiency test is aligned to the forms and functions of the Oregon ELP content standards and describes the English proficiency of students based on six domains: Total Proficiency, Listening, Speaking, Reading, Writing, and Comprehension. Comprehension is a combination of the Reading and Listening measures. Total Proficiency is a combination of Listening, Speaking, Reading, and Writing. The item types of speaking short response (SSR), speaking extended response (SER), and writing extended response (WER) are scored in two dimensions, grammar and illocution. Starting in 2011–12, the grammatical and illocutionary competence scores were also reported.

Oregon's ELP assessment is designed to satisfy the provisions of Title III of ESEA. Scores are to be used for the following:

- Providing an annual English language proficiency score and level for each student
- Reporting annual measures of speaking, reading, listening, writing, and comprehension for each student [(Section 3113(b)(3)(D)]
- Reporting Annual Measurable Achievement Objectives (AMAOs) biennially to the federal government (Because ELLs enter school systems at different ages with different degrees of English proficiency, AMAOs can be based on cohorts, groups of students entering at a common age and proficiency level.)

1.2 Oregon's English Language Proficiency (ELP) Standards

The Oregon Department of Education, in partnership with educators throughout the state, developed Oregon's English Language Proficiency Standards which were adopted by the State Board of Education in 2005. These standards describe progressive levels of competence in English acquisition for five proficiency levels: beginning, early intermediate, intermediate, early advanced, and advanced. English language proficiency levels set clear proficiency benchmarks of progress at various grade levels.

As specified in Title III of ESEA, ELP content standards are designed to supplement the existing ELA academic content standards to facilitate students' transitioning into regular education content classes. ELP standards were designed to guide language acquisition to allow English Language Learners to successfully participate in regular education classes. ELP assessments measure ELP standards, not ELA standards. This is an important distinction, as ELP content validity is based on the degree to which tests reflect ELP content standards, which, although designed to supplement the ELA standards, are quite different in structure and meaning. ELLs are required to take ELP assessments *in addition to* ELA and other content assessments. Therefore, the domain of ELP assessments differs from that of ELA.

1.3 Oregon's English Language Proficiency Assessment (ELPA)

The State of Oregon ELPA has the following features:

- Web-based adaptive
- Research-based and documented
- Aligned to the Oregon ELP (English Language Proficiency) standards
- Aligned to the Oregon ELA (English Language Arts) content standards
- Valid and reliable
- Conducted in English
- Tests the following grade bands: K–1, 2–3, 4–5, 6–8, and 9–12 and is required of all ELLs enrolled in these grades
- Produces a score and level for overall academic English proficiency (Cut points are established on the overall English proficiency scale.)
- Produces sub-scores in four domains: listening, speaking, writing, and reading
- Reports a measure of comprehension as a combination of listening and reading
- Reports a measure of grammar as a sum of grammar dimension scores from SSR, SER, and WER items
- Reports a measure of illocution as a sum of illocution dimension scores from SSR, SER, and WER items
- Demonstrates growth in English language acquisition skills over time.
- Applicable to students of any language or cultural background
- Supports Title I accountability and Title III program evaluation in local school districts

2. TEST DEVELOPMENT

Following the adoption and integration of the Oregon English language proficiency standards into the school curricula, item and test specifications (linked here) were developed to make sure that the tests and their items are aligned to the standards and grade-level expectations they are intended to measure. These item and test specifications identify the item types, quantity, and point values to be included in the assessments. These specifications also include the distribution of items across various content areas, such as mathematics and science. Specifications for reading tests include rules for identifying and selecting appropriate reading passages. The Oregon Department of Education, in partnership with educators throughout the state, developed Oregon's English Language Proficiency Standards. These standards describe progressive levels of competence in English acquisition for five proficiency levels: beginning, early intermediate, intermediate, early advanced, and advanced. English language proficiency levels set clear benchmarks of progress that reflect differences for students entering school at various grade levels.

The item development process consists of six major steps.

1. Review of the current item pool
2. Development of new items
3. Departmental item review and approval
4. Committee review of new items
5. Field-testing
6. Rangefinding and rubric validation by committee

These steps are detailed below.

2.1.1 Review of the Current Item Pool

Before item development begins, development targets are set based on a review of the existing item pool and in consultation with the Oregon Department of Education (ODE). Development targets are set for each domain, grade, item type, and proficiency level.

2.1.2 Development of New Items

The next required step is to develop items that measure the English Language Proficiency standards. Oregon teachers and/or testing contractor content specialists initially write the items to meet the development targets. The items then must pass through several internal review stages including content, editorial, and senior content reviews. The items are reviewed for proper alignment to the ELP standards, proficiency level, accuracy, grammatical correctness, and bias.

The Department returns reviews of the items with indications to approve, reject, or revise them. Recommendations for ways to improve the items are often included in the Department review.

Following the completion of the AIR and ODE internal item development cycle, ODE then convenes two committees, the Sensitivity Panel, which consists of Oregon teachers and community members, and the ELPA Content and Assessment Panel, which comprises Oregon teachers from across the state. The Sensitivity review ensures that the items remain free from bias or stereotype; the Content Panel review determines whether the items are properly aligned to the English language proficiency standards and grade-level expectations, and accurately measure intended content.

2.1.3 Departmental Item Review and Approval

After reviewing the items internally, the contractor sends batches of items to ODE for review. ODE reviews the items and provides an outcome (accept as appears, accept as revised, revise and resubmit, or reject) for each item. Recommendations for ways to improve the items are often included in the Department review. The testing contractor and ODE staff then discuss the requested revisions, ensuring that all items appropriately measure the Oregon English language proficiency standards. After discussion, the contractor makes requested revisions, and ODE approves the items that are ready for Oregon committee review.

2.1.4 Committee review of new items

All items generated for use on Oregon statewide assessments must pass a series of rigorous committee reviews before they can be used in field and operational tests. The items undergo two committee reviews during the development process: Sensitivity Panel review and the ELPA Content and Assessment Panel review.

The Sensitivity Panel reviews items for bias, controversial content, and emotionally sensitive issues. The Sensitivity Panel consists of Oregon educators and community members who are selected to ensure geographic and ethnic diversity. The committee ensures that items

- present racial, ethnic, and cultural groups in a positive light;
- do not contain controversial, offensive, or potentially upsetting content;
- avoid content familiar only to specific groups of students because of race or ethnicity, class, or geographic location;
- aid in the elimination of stereotypes; and
- avoid words or phrases that have multiple meanings.

ODE and the contractor reject or edit items based on Sensitivity Panel feedback.

The ELPA Content and Assessment Panel consists of Oregon educators familiar with English language proficiency and grades being reviewed. Committees are selected to ensure geographic and ethnic diversity. Items are accepted, rejected, or modified by the Content Panel members to make sure they accurately measure the Oregon ELP Standards. Only the items that measure these standards are carried forward to the field-test stage. In addition to ensuring standards alignment, the Content Panel members review the items to ensure that they are free from such flaws as (a) inappropriate readability level, (b) ambiguity, (c) incorrect or multiple answer keys, (d) unclear instructions, (e) misaligned proficiency level, and (f) factual inaccuracy.

2.1.5 Field-testing

Items that are approved by both panels and ODE are placed on field tests during the next administration. Scores for all item types are recorded. Human-scored constructed-response items including all SER, SSR, WER, and Elicited Imitation (EI) items are sent for review by the Rangefinding Committee. Technologically enhanced items such as the Wordbuilder items are sent for review to the Rubric Validation committee before final scores for those items are produced.

2.1.6 Rangefinding and Rubric validation by committee

The rubrics for the field-test machine-scored constructed-response (MSCR) items go through a validation process to refine the machine-scored rubrics. The rubric validation process is analogous to rangefinding for human-scored items, checking the validity of scoring rubrics as well as the scoring technology. The rubric validation process uses a committee of content area experts to review student responses and propose changes to the machine-scored rubric.

During rubric validation, English language proficiency educators review the machine assigned scores for every MSCR item and either approved the score or suggested a revised score based on their interpretation of the item task and the rubric. The committee reviews 45 responses for each question.

If any scores change based on the rubric validation committee review, contractor staff revises the machine rubric and rescores the item. After the items are rescored, contractor staff review at least 10% of responses for which the score changed to ensures that committee suggestions are honored, that the item scores consistently, and that no unintended changes in scoring occur as a result of the revision to the machine rubric. Contractor staff review changes with ODE staff, and ODE staff have one final opportunity to revise the rubric or approve or reject the item.

The rangefinding committee reviews the rubric of the handscored items. The committee reviewed 10–15 sample papers per score point. For a two point item, about 20–30 sampled responses were provided, and for a three point item, about 30–45 sampled responses were drawn. The committee reviewed these sampled papers and adjusted the rubrics as needed to obtain valid measure of the language constructs in the item. The committee has the option to approve the score or suggest a different score based on the committee's understanding of the rubric. ODE and contractor staff ensure that the committee scores in a consistent manner.

For items that have the grammatical and illocutionary dimensions, the committee reviews both dimensions for the same item. Items that survive the rangefinding or the rubric validation process are included in the operational pool.

In June 2014, 47 handscored field test items were reviewed by the rangefinding committee and another 172 field test grid items were reviewed by the rubric validations committee. After the committees' review, four field test items were rejected – one grade 2-3 WB item, two grade 4-5 grid items, and one grade 9-12 grid items. Another 34 field test items were put in "deferred" status – to be modified and refield-tested in the future. Twenty-two of these deferred items were the hot text type of items, a new item type developed for 2013–2014, eight grid items, two speaking short response items and two WB items.

3. Summary of Operational Procedures

3.1 Test Administration

The 2013–2014 ELPA was administered online during the testing window January 8, 2014 through April 15, 2014. ELPA tests were administered in two segments: segment one, made up of listening, reading, and writing (ELPA CORE); and Segment Two, the speaking segment, which was delivered to students after completing the Segment One. Field-test items were embedded during the entire testing window.

One testing opportunity was allowed per student for ELPA. Table 1 presents the number of students who participated in the 2013–2014 ELPA.

Table 1. Number of Students Participated in 2013–14 ELPA, by Grade Bands

Group	Grade K–1		Grade 2–3		Grade 4–5		Grade 6–8		Grade 9–12	
	N	%	N	%	N	%	N	%	N	%
All Students	16,106	100	14,642	100	11,090	100	6,980	100	5,028	100
Female	7,810	48	6,995	48	5,096	46	2,943	42	2,191	44
Male	8,296	52	7,647	52	5,994	54	4,037	58	2,837	56
African American	302	2	267	2	219	2	236	3	280	6
Asian American	1,460	9	1,334	9	1,043	9	787	11	980	19
Hispanic	12,595	78	11,561	79	8,811	79	5,304	76	3,276	65
White	1,441	9	1,175	8	800	7	494	7	408	8
Economically Disadvantaged	13,264	82	12,755	87	9,721	88	6,034	86	3,974	79

In 2013–2014, a student could be exempted from taking tests in one or more of the ELPA domains. Overall, there were 330 students exempted from one or more domains in 2013–2014, as compared to 216 exemptions of 2012–13. Table 2 presents a breakdown of the students who had exemptions by grade band. The categories with the largest exemptions were "Speaking Only" and "Reading and Writing."

Table 2. Number of Students Exempted in the 2013–2014 ELPA, by Domains and Grade Bands

Exemptions	Grade Bands					
	K–1	2–3	4–5	6–8	9–12	Total
None	16,047	14,576	10,988	6,892	5,013	53,516
Speaking Only	33	13	18	8	1	73
Writing Only	0	0	3	5	0	8
Writing and Speaking	1	1	1	4	2	9
Reading Only	2	4	8	16	0	30
Reading and Writing	6	26	64	43	6	145
Reading, Writing, and Speaking	11	14	6	4	2	37
Listening Only	0	1	0	0	2	3
Listening and Speaking	6	7	2	4	2	21
Listening, Writing, and Reading	0	0	0	1	0	1
Listening and Reading	0	0	0	3	0	3

3.1.1 Summary of Simulation Studies Performed Prior to the Operational Testing Window

Prior to the operational testing window, AIR conducts simulations to evaluate and ensure the implementation and quality of the adaptive item-selection algorithm and the scoring algorithm. The simulation tool enables us to manipulate key blueprint and configuration settings to match the blueprint and minimize measurement error and to maximize the number of different tests seen by students.

3.1.2 Testing Plan

Our testing plan begins by generating a sample of examinees from a normal (μ, σ) distribution for each grade band. The mean parameter for the normal distribution is taken from the 2012–2013 administration. Each simulated examinee is administered one opportunity. Because no prior information about the examinee is available, each student is assumed to have the same initial (theta) starting value. The starting value is used to initiate the test by choosing the first few items. The state average theta value in 2012–2013 was used as the starting theta.

Table 3 provides the means and standard deviations used in the simulation for each grade band.

Table 3. Mean and Standard Deviation Used in 2013–2014 Simulation

Grade Band	Mean	SD
K–1	–0.6	1.31
2–3	0.9	1.16
4–5	1.5	1.06
6–8	1.7	1.04
9–12	1.6	1.17

3.1.3 Statistical Summaries

The statistics computed include the statistical bias of the estimated theta parameter (statistical bias refers to whether test scores systematically underestimate or overestimate the student's true ability); mean squared error (MSE); significance of the bias; average standard error of the estimated theta; the standard error at the 5th, 25th, 75th, and 95th percentiles; and the percentage of students falling inside the 95% and 99% confidence intervals.

Computational details of each statistics are provided below.

$$bias = N^{-1}\sum_{i=1}^{N}(\theta - \hat{\theta}) \ (1)$$

$$\boldsymbol{MSE = N^{-1}\sum_{i=1}^{N}(\theta - \hat{\theta})^2}$$

where θ is the true score and $\hat{\theta}$ is the observed score. For the variance of the bias, we use a first-order Taylor series of Equation (1) as:

$$\mathrm{var}(\boldsymbol{bias}) = \sigma^2 * \boldsymbol{g}'(\hat{\theta})^2$$
$$= \frac{1}{\boldsymbol{N}(\boldsymbol{N}-1)}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}(\theta_{\boldsymbol{i}} - \bar{\hat{\theta}})^2$$

Significance of the bias is then tested as:

$$\boldsymbol{z} = \boldsymbol{bias} / \sqrt{\mathrm{var}(\boldsymbol{bias})} \, .$$

A *p*-value for the significance of the bias is reported from this *z* test.

The average standard error is computed as:

$$mean(se) = \sqrt{N^{-1}\sum_{i=1}^{N} se_i^2}$$

where se_i is the standard error of the estimated *θ* for individual *i*.

To determine the number of students falling outside the 95% and 99% confidence interval coverage, a *t*-test is performed as follows:

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)}$$

where $\hat{\theta}$ is the ability estimate for individual i, and θ is the true score for individual i. The percentage of students falling outside the coverage is determined by comparing the absolute value of the t-statistic to a critical value of 1.96 for the 95% coverage and to 2.58 for the 99% coverage.

3.2 Summary of Statistical Analyses

3.2 .1 Summary Statistics on Test Blueprints

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. All simulated tests met the test blueprint 100% in all grade bands for all four domains: reading, writing, listening and speaking.

3.2.2 Summary Bias and Standard Errors

Table 4 provides a summary of the bias and average standard errors of the estimated theta by grade band. The mean bias of the estimated abilities is larger in grade bands 4–5 and 6–8 than the bias in the other grade band, s The mean bias of 2013-14 is similar to that of 2012-13.

The summary statistics of the estimated abilities show that for all examinees in all grade bands, the item selection algorithm is choosing items that are optimized conditioned on each examinee's ability. Essentially, this shows that the examinee-ability estimates generated on the basis of the items chosen are optimal in the sense that the final score for each examinee always recovers the true score within expected statistical limits. In other words, given that we know the true score for each examinee in a simulation, these data show that the true score is virtually always recovered—an indication that the algorithm is working exactly as expected for a computer-adaptive test.

In addition, the average standard errors ranged from .28 to .31across all grade bands. With the exception of K-1, the average standard errors of the estimated abilities are similar across the ability ranges with the largest standard error at the 95th percentile, indicating a shortage of difficult items to better match the high-ability students. For K-1, the standard error in the 5th percentile is larger than the standard error at the 95th percentile. Although efforts have been made to augment the item pool with difficult items to measure the high-performing students' ability more efficiently, it is very challenging to develop item pools that are robust enough to accurately measure students at the extreme levels of knowledge and skills.

Overall, these diagnostics on the item-selection algorithm provide evidence that scores are comparable with respect to the targeted content and scores at various ranges of the score distribution are measured with good precision.

Table 4. Standard Errors of the Estimated Abilities (from simulation results) by Grade Bands, 2013–2014

Grade Band	Bias	Average Standard Error	SE at 5 Percentile	SE at Bottom Quartile	SE at Top Quartile	SE at 95 Percentile
K–1	-0.03	0.31	0.35	0.29	0.29	0.33
2–3	0.05	0.28	0.27	0.27	0.28	0.35
4–5	0.12	0.29	0.25	0.25	0.31	0.39
6–8	0.10	0.28	0.26	0.25	0.29	0.40
9–12	0.07	0.28	0.26	0.25	0.29	0.39

3.3 Summary of Adaptive Algorithm

For ELPA, item selection rules ensure that each student receives a test representing an adequate sample of the domain with appropriate difficulty. The algorithm maximizes the information for each student and allows for certain constraints to be set, ensuring that items selected represent the required content distribution. Items selected for each student depend on the student's performance on previously selected items. The accuracy of the student responses to items determines the next items and passage that the student will see. Thus, each student is presented with a set of items that most accurately align with his or her proficiency level. Higher performance is followed by more difficult items, and lower performance is followed by less difficult items until test length constraints are met. Because no prior information about the examinee is available, each student is assumed to have the same initial (theta) starting value. The starting value is used to initiate the test by choosing the first few items.

After the initial item is administered, the algorithm identifies the best item to administer using the following criteria:

3.3.1 MATCH TO THE BLUEPRINT

The algorithm first selects items to maximize fit to the test blueprint. Blueprints specify a range of items to be administered in each domain for each test, with a collection of *constraint sets*. A constraint set is a set of exhaustive, mutually exclusive classifications of items. For example, if a content area consists of four content domains and each item measures one—and only one—of the domains, the content domain classifications constitute a constraint set.

During item selection, the algorithm "rewards" domains that have not yet reached the minimum number of items. For example, if the listening content domain requires that a test contain between 10 and 15 items, listening is the constrained feature. At any point in time, the minimum constraint on some features may have already been satisfied, while others may not have been. Other features may be approaching the maximum defined by the constraint. The value measure must reward items that have not yet met minimum constraints and penalize items that would exceed the maximum constraints. The algorithm stops administering items when the specified test length is met.

3.3.2 Match to student ability

In addition to rewarding items that match the blueprint, the adaptive algorithm also places greater value on items that maximize test information near the student's estimated ability, ensuring the most precise estimate of student ability possible, given the constraints of the item pool and satisfaction of the blueprint match requirement. After each response is submitted, the algorithm recalculates a score. As more answers are provided, the estimate becomes more precise, and the difficulty of the items selected for administration more closely aligns to the student's ability level. Higher performance (answering items correctly) is followed by more difficult items, and lower performance (answering items incorrectly) is followed by less difficult items. When the test is completed, the algorithm scores the overall test and each content domain.

The algorithm allows previously answered items to be changed; however, it does not allow items to be skipped. Item selection requires iteratively updating the estimate of the overall and domain ability estimates after each item is answered. When a previously answered item is changed, the proficiency estimate is adjusted to account for the changed responses when the next new item is selected. While the update of the ability estimates is performed at each iteration, the overall and domain scores are recalculated using all data at the end of the test for the final score.

For the speaking item selection, stakeholders raised a concern that for those students who may be more proficient in speaking than in the other domains (e.g., listening, reading, and writing), the selection of speaking items on such a student's performance in the other domains might artificially lower the difficulty of the speaking items presented to the student and therefore prevent the student from accurately demonstrating proficiency in speaking.

The speaking segment was delivered to students after completing Segment one, made up of listening, reading, and writing. Based on results of analyses, a semi-adaptive approach was taken. The speaking items delivered to a student in Segment two were derived from both (1) a group of items on which the student may demonstrate advanced proficiency and (2) items selected based on the scaled score the student achieved on Segment one of the ELPA.

The ELPA speaking test blueprint was revised to ensure that each student receives at least one speaking item with a sufficiently high scale value so that each student has the opportunity to demonstrate advanced speaking skills regardless of his or her performance on the other domains. The rubrics for the speaking short response (SSR) and the speaking extended response (SER) items are multiple point rubrics, so that students at all proficiency levels have the opportunity to earn points on the item. Even students at the beginning level have the opportunity to score points because of these multi-point rubrics, while advanced speakers can score maximum points.

The semi-adaptive selection approach to speaking items is described below for each grade band.

Grades K–1 Blueprint: Nine Speaking Short-Response Items and 19 Elicited Imitation Items available

Each kindergarten and 1st grade student is presented with eight speaking items (elicited imitation and short response). The first two speaking items presented to a student are determined on the basis of his or her performance on the machine-scored listening, writing, and reading items. Each student is then presented with at least one and up to three items in the highest difficulty range for the K–1 grade band. The purpose of these difficult items is to give each student an opportunity to demonstrate their speaking skills at a more advanced level regardless of his or her score on the machine-scored domains (reading, writing, and listening).

Grades 2–3 Blueprint: Seven Speaking Extended-Response Item, Six Speaking Short-Response Item, and 12 Elicited Imitation Items available

Each 2nd and 3rd grade student is presented with six speaking items (extended response, short response, and elicited imitation). The first two speaking items presented to a student are determined on the basis of his or her performance on the machine-scored listening, writing, and reading items. Each student is then presented with at least one and up to four items in the highest difficulty range for the 2–3 grade band. The purpose of these difficult items is to give each student an opportunity to demonstrate speaking skills at the advanced level regardless of his or her score on the machine-scored domains (reading, writing, and listening).

Grades 4–5 Blueprint: Twelve Speaking Extended-Response Items and Six Speaking Short-Response Item available

Each 4th and 5th grade student is presented with three speaking items (extended response and short response). The first two speaking items presented to a student are determined on the basis of his or her performance on the machine-scored listening, writing, and reading items. Each student is then presented with one and up to three items in the highest difficulty range for the 4–5 grade band. The purpose of these difficult items is to give each student a chance to demonstrate speaking skills at the advanced level regardless of his or her score on the machine-scored (reading, writing, and listening) domains.

Grades 6, 7, and 8 Blueprint: Twenty-Three Speaking Extended-Response Items available

Each 6th, 7th, and 8th grade student is presented with three speaking items (extended response). The first two speaking items presented to a student are determined on the basis of his or her performance on the machine-scored listening, writing, and reading items. Each student is then presented with one and up to three items in the highest difficulty range for the 6–8 grade band. The purpose of these three items is to give each student an opportunity to demonstrate speaking skills at the advanced level regardless of his or her score on the machine-scored domains (reading, writing, and listening).

High School Blueprint: Twenty-Three Speaking Extended-Response Items

Each student in grades 9–12 is presented with three speaking items (extended response). The first two speaking items presented to a student are determined on the basis of his or her performance on the machine-scored listening, writing, and reading items. Each student is then presented with one and up to three items in the highest difficulty range for the high school grade band. The purpose of these three items is to give each student an opportunity to demonstrate speaking skills at the advanced level regardless of his or her score on the machine-scored domains (reading, writing, and listening).

4. FIELD-TESTING

4.1 Administration

Oregon uses an embedded "operational" field-test design to augment items in the item pool. Each operational test embeds three to five field-test items for non-speaking items and one field-test item for speaking.

In each year, the field-test item development plan is based on a thorough review of the current operational item pool. The field-test items are developed to increase the items covering a full range of item difficulties and item types in need of improvement. In 2013–2014, 316 field-test items were administered, and these embedded field-test items were not counted toward student scores. Thirty-eight of the field-test items measured both grammatical and illocutionary competence, generating a total of 354 field-test parameters. Table 5 presents the number of field-test items embedded in 2013–2014 by item type and grade band. Twenty-six hot text items, is a new item type developed for 2013–2014, are included as MSCR items in Table 5. The number of ELPA test participants of 2012–13 was used to estimate the number of field test items can be embedded for each grade band so that each field test item will have at least 550 responses without significantly increases the overall test length.

Table 5. Number of Field-Test Items by Item Type and Grade Bands, ELPA 2013–2014

	Item Type						
Grade Band	Elicited Imitation	MSCR Items	Multiple Choice	Short Response	Speaking Extended Response	Word Builder	Total
K–1	5	24	34	6		5	74
2–3	4	27	21	3	4	9	68
4–5		36	17	5	5		63
6–8		38	15	4	4		61
9–12		33	10	3	4		50
All Grades	9	158	97	21	17	14	316

The ELPA item pool includes six types of items: multiple-choice items, machine-scored constructed response items (MSCR), word builder (WB), elicited imitation (EI), speaking short response (SSR), speaking extended response (SER), and writing extended response (WER). Multiple choice, MSCR and WB items are machine-scored. EI, SSR, and ER items are human-scored.

4.2 Sample Selection and Item Administration Selection Algorithm

AIR's field-test sampling algorithm is designed to yield an efficient, scientifically sound, representative random sample. The field-test item administered to a student is selected randomly from among those that have been *least frequently administered*, and this produces a similar sample size for all items with subgroup compositions similar to the population for each item. This is a very powerful sample design that will yield a representative sample.

The algorithm employed by AIR's field-test engine ensures that

- efficient samples are used for all items by randomly selecting from among the least frequently administered items to ensure that resulting item statistics (and DIF statistics) are maximally efficient across all items;
- position effects are averaged out by randomly administering test items across test positions; and
- more robust linkages exist among items, because each item is linked with every other item across hundreds of unique test "forms" to a degree not manageable through a set of fixed forms.

For pragmatic considerations related to system performance, the field-test algorithm limits the item selection to test start-up, instead of selecting each item in real time as with the adaptive algorithm. Therefore, the field-test engine assigns all items at the beginning of each test administration.

Upon test start-up, the algorithm selects a series of items in the following iterative sequence:

1. Identify all the items that were least frequently administered
2. Randomly select an item with equal probability
3. Return to step 1 and continue if the requisite number of items has not been met

The first step initiates a random sequence. Note that for the first student, all items may be selected with equal probability. Subsequent selections depend on this first selection because the item is sampled without replacement until the entire pool has been administered, at which point the whole set becomes eligible for sampling again. This dependence is analogous to the dependence that occurs when test books are spiraled within classrooms—the probability of a student being assigned a particular book depends on the availability of books left to be handed out. In both cases the temporal dimension is incidental: the end result is a random distribution of items (or books) within the classroom.

We can see that the probability of administering an item from the pool is constant across individuals in the population. For a single grade and content area, let n represent the number of students, k represent the number of field-test items on the test, and m represent the total number of items in a pool. The expected number of times that any single item j will be administered can be calculated by $n_j = \frac{nk}{m}$. The corresponding probability that a given student i will receive item j is therefore $p_i(j) = \frac{n_j}{n} = \frac{k}{m}$.

From this we see that

- a random sample of students receives each item;
- for any given item, the students are sampled with equal probability.

This design is both randomized, ensuring representation of the population and the validity of estimates of sampling error, and efficient.

The field-test algorithm also leads to randomization of item position and the context in which items appear. Field-testing each item in many positions and contexts should render the resulting statistics more robust to these factors.

4.3 Field-Test Item Position

For 2013–2014, field-test items in all grade bands are positioned throughout the test, avoiding the beginning or end of the operational test. Each operational test embeds three to five non-speaking field-test items and one speaking field-test item.

Table 6 presents the number of field test slots and the field-test item positions in each grade band.

Table 6. Targeted Number of Embedded Field-Test (FT) Items and Item Position, by Segments and Grade Bands, ELPA, 2013–2014

Segment	Grade Band	FT Min. Items	FT Max. Items	FT Starting Position	FT Ending Position
CORE	K–1	3	3	5	35
	2–3	3	5	5	45
	4–5	3	5	5	45
	6–8	5	5	5	45
	9–12	5	5	5	45
Speaking	K–1	1	1	4	5
	2–3	1	1	4	5
	4–5	1	1	2	3
	6–8	1	1	2	3
	9–12	1	1	2	3

5. Embedded Field-Test Item Analyses Overview

5.1 Field-Test Item Analyses

Once the scoring rubrics for all machine-scored constructed-response items are validated, all constructed-response items are rescored using the final rubrics, and the final data file is extracted for the item analyses. The item analyses include classical item statistics and item calibrations using the Rasch-family IRT models. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (DIF analyses). Sample size for field-tested items ranged from 588 to 1,463, across grade bands.

5.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item is able to differentiate between high- and low-achieving students. The discrimination index is calculated as the correlation between the item score and the student's IRT-based ability estimate (biserial correlations for multiple-choice items and polyserial correlations for constructed-response items). Multiple-choice items are flagged for subsequent reviews if the correlation for the item is less than .20 for the keyed (correct) response and greater than .05 for distractors. For constructed-response items, items are flagged if the polyserial correlation is less than .20.

5.1.2 Item Difficulty

Items that are either extremely difficult or extremely easy are flagged for review but are not necessarily rejected if the item discrimination index is not flagged. For multiple-choice items, the proportion of examinees in the sample selecting the correct answer (*p*-values), as well as those selecting incorrect responses, is computed. For constructed-response items, item difficulty is calculated both as the item's mean score and as the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score divided by the number of points possible). Items are flagged for review if the *p*-value is less than .25 or greater than .95.

Constructed-response items are flagged if the proportion of students in any score-point category is greater than .95. A very high proportion of students in any single score-point category may suggest that the other score points are not useful or, if the score point is in the minimum or maximum score-point category, that the item may not be grade-appropriate. Constructed-response items are also flagged if the average IRT-based ability estimate of students in a score-point category is lower than the average IRT-based ability estimate of students in the next lower score-point category. For example, if students who receive three points on a constructed-response item score, on average, lower on the total test than students who receive only two points on the item, the item is flagged. This situation may indicate that the scoring rubric is flawed.

Table 7 below provides a five-point summary of the *p*-values by grade band and range. The min. column lists by grade the smallest *p*-values, and the columns to the right represent the *p*-value at a particular percentile—the 5th percentile, the 25th percentile, and so on.

Table 7. Summary of p-Value Range for 2013–2014 Field Test Items, by Grade Bands

Grade Band	p5	p25	p50	p75	p95
K–1	0.19	0.44	0.55	0.78	0.89
2–3	0.25	0.5	0.65	0.77	0.92
4–5	0.5	0.77	0.835	0.9	0.95
6–8	0.32	0.58	0.75	0.83	0.95
9–12	0.35	0.465	0.6	0.755	0.84

5.1.3 Differential Item Functioning (DIF)

DIF analyses are designed to determine whether students at similar levels of ability have different probabilities of answering the same item correctly (or of receiving higher scores in the case of constructed-response items), based on a group membership. In some cases, DIF may indicate item bias. However, a subcommittee of the ELPA Content and Assessment Panel will review items classified as DIF to determine whether an item is unfair to members of various student subgroup populations.

AIR conducted DIF analyses on all field-test items to detect potential item bias for subgroups. DIF analyses were performed for the following groups in Oregon:

- Male/female
- Hispanic/African American
- Hispanic/Asian American
- White/African American
- White/Asian American
- White/Hispanic
- Economically disadvantaged/not economically disadvantaged

Differential item functioning refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF.

AIR typically uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. IRT ability estimates for each student on the test are used as the ability-matching variable. That score is divided into five intervals to compute the MH chi-square DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta for the multiple-choice items, the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the constructed-response items. The purification method described by Holland and Thayer (1988) is included in the DIF procedure.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention illustrated in Table 7. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. These items are flagged regardless of whether the DIF statistic favors the focal or reference group. ELPA Content and Assessment Panel members review the items flagged on the basis of DIF statistics. Committee members are encouraged to discuss these items and are asked to decide whether each item should be excluded from the pool of potential items given its performance in field-testing. Table 8 details the DIF classification rules. Results from the DIF analysis for the non-rejected and non-deferred field test items are presented in Table 9. In this table, each dimension of an item is counted separately.

Table 8. DIF Classification Rules

DIF Category	Flag Criteria
Dichotomous Items	
C	$MH\chi^2$ is significant and $\mid \hat{\Delta}_{MH} \mid \geq 1.5$.
B	$MH\chi^2$ is significant and $\mid \hat{\Delta}_{MH} \mid < 1.5$.
A	$MH\chi^2$ is not significant.
Polytomous Items	
C	$MH\chi^2$ is significant and $\mid SMD \mid / \mid SD \mid \geq .25$.
B	$MH\chi^2$ is significant and $\mid SMD \mid / \mid SD \mid < .25$.
A	$MH\chi^2$ is not significant.

Table 9. Results from DIF Analysis of the 2013–2014 Field-Test Items

DIF Groups (Focus/Reference)	Categories		Grade Bands				
			K–1	2–3	4–5	6–8	9–12
Disadvantaged/Non-Disadvantaged	A	+	35	31	28	29	22
		–	39	32	21	30	20
	B	+	1	2	4	1	5
		–	3	.	.	1	1
	C	+	1	2	5	1	1
		–	.	.	.	.	1
Black/Hispanic	A	+	38	30	16	27	17
		–	37	30	30	31	26
	B	+	.	.	.	.	.
		–	.	.	1	.	2
	C	+	2	2	.	.	2
		–	2	5	11	4	3
Asian/Hispanic	A	+	38	35	16	28	19
		–	33	24	34	25	14
	B	+	.	.	.	1	5
		–	1	5	4	4	4
	C	+	4	1	1	.	2
		–	3	2	3	4	6
Female/Male	A	+	40	34	25	28	25
		–	27	22	26	24	20
	B	+	6	8	4	2	3
		–	4	3	2	5	2
	C	+	1	.	1	.	.
		–	1	.	.	3	.
Black/White	A	+	42	31	22	25	20
		–	32	31	30	34	24
	B	+	.	.	.	.	1
		–	.	.	1	.	.
	C	+	1	2	.	2	3
		–	4	3	4	1	2

DIF Groups			Grade Bands				
(Focus/Reference)	Categories		K–1	2–3	4–5	6–8	9–12
Asian/White	A	+	34	31	30	19	24
		–	39	31	24	39	22
	B	+	.	.	.	.	1
		–	1	.	.	1	.
	C	+	3	4	1	1	1
		–	2	1	3	2	2
Hispanic/White	A	+	30	36	32	23	25
		–	39	28	17	32	20
	B	+	1	.	2	2	1
		–	3	.	1	1	.
	C	+	4	3	5	2	3
		–	2	.	1	2	1

5.2 Item Data Review Committee Meetings

Items flagged for review on the basis of any of the aforementioned criteria must pass a three-stage data review process to be included in the final item pool from which operational forms are created. As a first level of review, a team of testing contractor psychometricians reviews all flagged items to ensure that the data are accurate, properly analyzed, have correct response keys, and have no obvious problems.

Second, the ODE ELPA specialist reviews the item statistics and content appropriateness of all flagged items. Then, a subcommittee of the ELPA Content and Assessment Panel reviews all items flagged on the basis of DIF statistics and other psychometric criteria. Panel members are encouraged to discuss these items using the statistics as a guide and are asked to decide whether flagged items should be excluded from the pool of potential items given their performance in field-testing.

6. ITEM CALIBRATION AND SCALING

6.1 Methodology

The ELPA items are calibrated using the one-parameter Rasch model (Rasch, 1980; Wright & Stone, 1979) for multiple-choice items and the Rasch partial-credit model (Masters, 1982) for constructed-response items, scored polytomously. Calibrating mixed item types from different assessment modes (i.e., dichotomously and polytomously scored items) requires the use of a polytomous model, which allows the number of score categories (typically score points on a scoring rubric) to vary across assessment modes. The Rasch partial credit model (Wright & Masters, 1982) can accommodate the mixing of dichotomous and polytomous items.

Under the Rasch model, the probability of a correct response conditional on ability is

$$\mathrm{p}(x_i = 1|\theta) = \frac{1}{1 + \exp[-(\theta - b_i)]} \quad (1)$$

where b_i is the location or difficulty parameter for the *i*th item, and x_i is the binary reponse to the *i*th item (where 1 = correct). The generalization for polytomous items in the partial credit model is

$$\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\exp \sum_{j=0}^{x}(\theta - \delta_{ij})}{\sum_{r=0}^{M}\left[\exp \sum_{j=0}^{r}(\theta - \delta_{ij})\right]} \quad (2)$$

where the notation is the same as Equation (1) other than δ_{ij}, which is the *j*th step for the *i*th item. Note that in the case of a dichotomous response item, Masters' model reduces to the Rasch model.

6.1.1 ITEM PARAMETER ESTIMATION

The Winsteps software program (Linacre, 2011) is used in the item calibration. Winsteps employs a joint maximum likelihood approach to estimation (JML), which estimates the item and person parameters simultaneously.

Given a series of responses to test items and the traditional assumption of conditional independence, the following likelihood function can be developed to express the joint likelihood of the item and ability parameters as:

$$\boldsymbol{L(\beta, \theta|x)} =$$
$$\prod_{i=1}^{N}\prod_{s=1}^{K}\left[\frac{\mathbf{1}}{\mathbf{1}+\exp[-(\boldsymbol{\theta_s}-\boldsymbol{b_i})]}\right]^{x_i}\left[\mathbf{1}+\frac{\mathbf{1}}{\mathbf{1}+\exp[-(\boldsymbol{\theta_s}-\boldsymbol{b_i})]}\right]^{\mathbf{1}-x_i}\prod_{i=1}^{N}\prod_{s=1}^{K}\frac{\exp\sum_{k=1}^{x_i}(\boldsymbol{\theta_s}-\boldsymbol{\delta_{ki}})}{\mathbf{1}+\sum_{j=1}^{m_i}\exp\sum_{k=1}^{j}(\boldsymbol{\theta_s}-\boldsymbol{\delta_{ki}})}.$$

where $\mathrm{b_i}$ is the location parameter for item i, $\mathrm{x_i}$ is the observed response to the item, θ_s is the ability of student s, and δ_{ki} is the kth step for item i with m total categories.

This likelihood function is composed of two latent parameters: item parameters, b_i and δ_{ki}, and student ability parameters, θ_s. The JML algorithm proceeds iteratively as follows (using β to denote item parameters for simplicity of notation):

1. Set $\beta = 0 \; \forall \; \mathrm{i}$ (or initialize to better starting values).
2. Set $\frac{\ln \partial \mathrm{L}(\beta)}{\partial \theta} = 0$ and solve.
3. Set $\frac{\ln \partial \mathrm{L}(\beta)}{\partial \beta} = 0$ and solve.
4. Iterate between steps 2 and 3 until the following two conditions hold:
   a. $\mathrm{abs}\left|\beta_j^t - \beta_j^{t-1}\right| <$ con $\forall \; j$, and
   b. $\mathrm{abs}\left|\theta_\mathrm{s}^\mathrm{t} - \theta_\mathrm{s}^{\mathrm{t}-1}\right| <$ con $\forall \; s$,

where *con* is a prespecified convergence criterion and the superscript denotes iteration *t*.

JML estimates are known to be asymptotically inconsistent. Consequently, we apply the traditional correction for bias in Winsteps as:

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} * (N-1)/N,$$

where *N* is the number of test items.

All items in the initial ELPA item pool, across grade bands, were concurrently calibrated, thus placing all test items on the same vertical scale. The embedded field-test items were concurrently calibrated fixing the precalibrated operational item parameters, placing the field-test items on the operational scale.

For tests that have domain exemptions, only items from the non-exempted domains are included in the likelihood function. Operational items in the exempted domains are treated as not-presented.

6.1.2 Rasch Descriptive Statistics

Table 10 below provides the five-point summary and range of the Rasch item parameters by grade band**.** The five-point difficulties indicate 5th percentile, 25th percentile, 50th percentile, 75th percentile, and 95th percentile of the Rasch item parameters.

Table 10. Rasch Item Parameter Five-Point Summary and Range, 2013–2014

Grade	Min	p5	p25	p50	p75	p95	Max
K–1	-3.58	-3.00	-2.09	-0.80	-0.21	1.30	2.57
2–3	-3.94	-2.02	-0.65	0.07	0.98	2.28	3.27
4–5	-2.26	-1.86	-1.03	-0.32	0.16	1.51	2.21
6–8	-2.18	-1.73	-0.07	0.46	1.30	2.53	3.08
9–12	-0.61	-0.56	0.11	1.13	1.64	2.22	2.36

.

6.1.3 Item Fit Index

The model fit is examined using the infit and outfit statistics. The infit statistic is more sensitive to the overall pattern of responses, less influenced by outliers, and more sensitive to patterns of observations by students on items that are roughly targeted for them. The outfit statistic is highly influenced by a few outliers (very unexpected observations) by students on items that are relatively very easy or very hard for them.

6.2 ELPA Scale

6.2.1 Overall Performance

The student’s performance is summarized in an overall test score referred to as a *scaled score*. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scaled scores so that scores from different sets of items can be meaningfully compared. The scaled scores represent a linear transformation of the Rasch ability estimates (theta scores), $SS=\hat{\theta}\ *\ Slope + Intercept,$ applying an intercept and a slope as following:

$$Scale\ Score = 500 + 10 * \hat{\theta}$$

where the theta $(\hat{\theta})$ represents any level of student ability on the operational form.

Standard errors of the ability estimates are also transformed to be placed onto the same reporting scale. This transformation is

$$se(SS) = 10 * se(\hat{\theta}).$$

The scaled scores are mapped into five performance levels using four performance standards (i.e., cut scores). In 2013–2014, ODE adopted a new set of cut scores for each of the performance levels. Table 11 provides the scaled score range used in 2013–2014 for each performance level for each grade. Table 11a shows changes in the scaled score at the lower end of the performance levels when comparing 2013–2014 with the 2012–13 cut scores. For example, Table 11a shows that in grade 1, the cut score at the lower end of the Early Intermediate level in 2013–2014 was 1 point lower (-1) than the 2012–13 cut score. The largest change is in the Advanced level for grade 12 students.

Table 11. Performance Standards for ELPA Scale Scores, 2013–2014

Grade	Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
K	430–480	481–490	491–496	497–504	505–570
1	430–490	491–502	503–511	512–521	522–570
2	430–491	492–503	504–513	514–520	521–570
3	430–499	500–510	511–520	521–525	526–570
4	430–493	494–503	504–513	514–521	522–570
5	430–495	496–507	508–514	515–523	524–570
6	430–492	493–503	504–515	516–521	522–570
7	430–494	495–507	508–517	518–523	524–570
8	430–496	497–508	509–519	520–526	527–570
9	430–493	494–499	500–512	513–522	523–570
10	430–493	494–499	500–512	513–522	523–570
11	430–493	494–499	500–512	513–522	523–570
12	430–493	494–499	500–512	513–522	523–570

Table 11a. Changes in the Scaled Cut scores at the Lower End of the Performance Level between 2012–13 and 2013–14

Grade	Early Intermediate	Intermediate	Early Advanced	Advanced
K	-1	-1	-1	-2
1	-1	-4	-2	-1
2	-3	-4	0	-2
3	-1	-3	0	-3
4	-3	-4	0	1
5	-1	0	-1	1
6	-4	-2	1	0
7	-2	1	1	0
8	-2	1	2	1
9	3	-1	-2	-3
10	1	-1	-3	-4
11	0	-1	-2	-5
12	-4	-4	-3	-7

.

6.2.2 Reporting Category Performance

In addition to the overall scaled score and performance levels, students receive a scaled score and a performance level for each reporting category. These scores are derived from using items that belong to that reporting category. That is, the likelihood function is maximized using only the subset of items measuring that reporting category. The same linear transformation of the Rasch ability estimates and the performance standards for the overall scaled scores and performance levels are applied to produce the scaled scores and performance levels for reporting category scores.

7. Statistical Summary of the Content Distribution and Measurement Characteristics of the Tests Delivered

7.1 Test Information and Standard Error of Measurement

In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the item response theory (IRT) framework, measurement error varies conditioned on ability. The amount of precision in estimating achievement can be determined by the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is the inverse of the measurement error of the test: the larger the measurement error, the less test information is being provided. In computer-adaptive testing (CAT), items vary across students, and the measurement error can vary for the same ability depending on the selected items for each student.

7.1.1 Marginal Reliability

For the reliability coefficient index, the *marginal reliability coefficients* were computed for the scaled scores, taking into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the ability scale, for all students.

The *marginal* reliability is defined as$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^{N} CSEM_i^2}{N}\right)]/\sigma^2$, where N is the number of students; $CSEM_i$ is the conditional standard error of measurement (SEM) of the scaled score of student *i;* and σ^2 is the variance of the scaled score. Standard error of measurement can be computed as $SEM = \sigma\sqrt{1-\bar{\rho}} = \sqrt{\sum_{i=1}^{N} CSEM_i^2 / N}$.

Table 12 presents the *marginal* reliability coefficients and the average standard error of measurements for the total scaled scores and reporting category scores. The marginal reliability coefficients for the total scaled score are similar across grade bands, ranging from .92 to .94. Coefficients are smaller for the reporting categories of listening, reading, writing and speaking, ranging from .63 in the listening subscale to .89 in the Writing subscale. The grammatical and illocutionary reporting categories in grade bands K–1 and 2–3 includes two items only, resulting in a negative reliability coefficient and a large standard error of measurement. The large standard error of measurements indicates that the grammar and illocution reporting categories in K–1 and 2–3 include too few items to report reliable scores.

The average conditional standard error of measurements at the threshold for each performance standard is also reported in Table 13. The average is largest in the advanced performance level, which can be expected given a shortage of very difficult items in this item pool, which measures high-performing students.

Table 12. Marginal Reliability and Standard Errors of Measurement for Reporting Categories (Content Domains), 2013–2014

Grade	Content Domain	Number of Items Specified in Test Blueprint		Marginal Reliability	Average Scaled Score	SD	SEM
		Min	Max				
K–1	Total Test	48	48	0.94	495	12.45	3.05
	Reading	10	15	0.81	494	14.51	6.39
	Listening	10	15	0.75	497	14.66	7.39
	Speaking	8	8	0.79	495	16.95	7.79
	Writing	14	19	0.85	495	15.22	5.91
	Comprehension	24	24	0.88	495	12.85	4.53
	Grammatical*	2	2	0.23	498	15.20	13.30
	Illocutionary*	2	2	0.26	499	15.57	13.38
2–3	Total Test	56	56	0.94	509	12.15	2.90
	Reading	10	15	0.82	509	15.20	6.53
	Listening	10	15	0.63	510	12.48	7.61
	Speaking	6	6	0.77	509	16.42	7.93
	Writing	23	32	0.89	509	14.02	4.56
	Comprehension	24	25	0.86	509	12.77	4.81
	Grammatical*	2	3	-0.08	507	12.83	13.35
	Illocutionary*	2	3	0.28	510	15.95	13.50
4–5	Total Test	53	53	0.92	515	10.51	2.92
	Reading	14	20	0.79	514	11.09	5.06
	Listening	14	20	0.67	515	11.51	6.65
	Speaking	3	3	0.66	518	16.75	9.76
	Writing	12	14	0.82	518	15.14	6.47
	Comprehension	36	37	0.86	514	10.14	3.83
	Grammatical	7	7	0.73	521	18.08	9.40
	Illocutionary	7	7	0.70	516	14.13	7.74
6–8	Total Test	53	53	0.93	517	10.80	2.82
	Reading	14	20	0.81	515	12.15	5.36
	Listening	14	20	0.75	516	11.02	5.53
	Speaking	3	3	0.71	518	16.60	8.92
	Writing	12	14	0.83	521	15.05	6.22
	Comprehension	36	37	0.87	516	10.38	3.72
	Grammatical	7	7	0.69	524	17.50	9.75
	Illocutionary	7	7	0.75	518	14.64	7.32
9–12	Total Test	53	53	0.94	515	11.80	2.82
	Reading	14	20	0.79	515	11.74	5.34
	Listening	14	20	0.79	514	12.64	5.79
	Speaking	3	3	0.76	518	18.57	9.12
	Writing	12	14	0.83	517	15.53	6.44
	Comprehension	36	37	0.89	514	11.25	3.79
	Grammatical	7	7	0.70	519	18.34	10.05
	Illocutionary	7	7	0.77	516	15.86	7.60

*Due to the small number of items and total points, the statistics displayed is unstable.

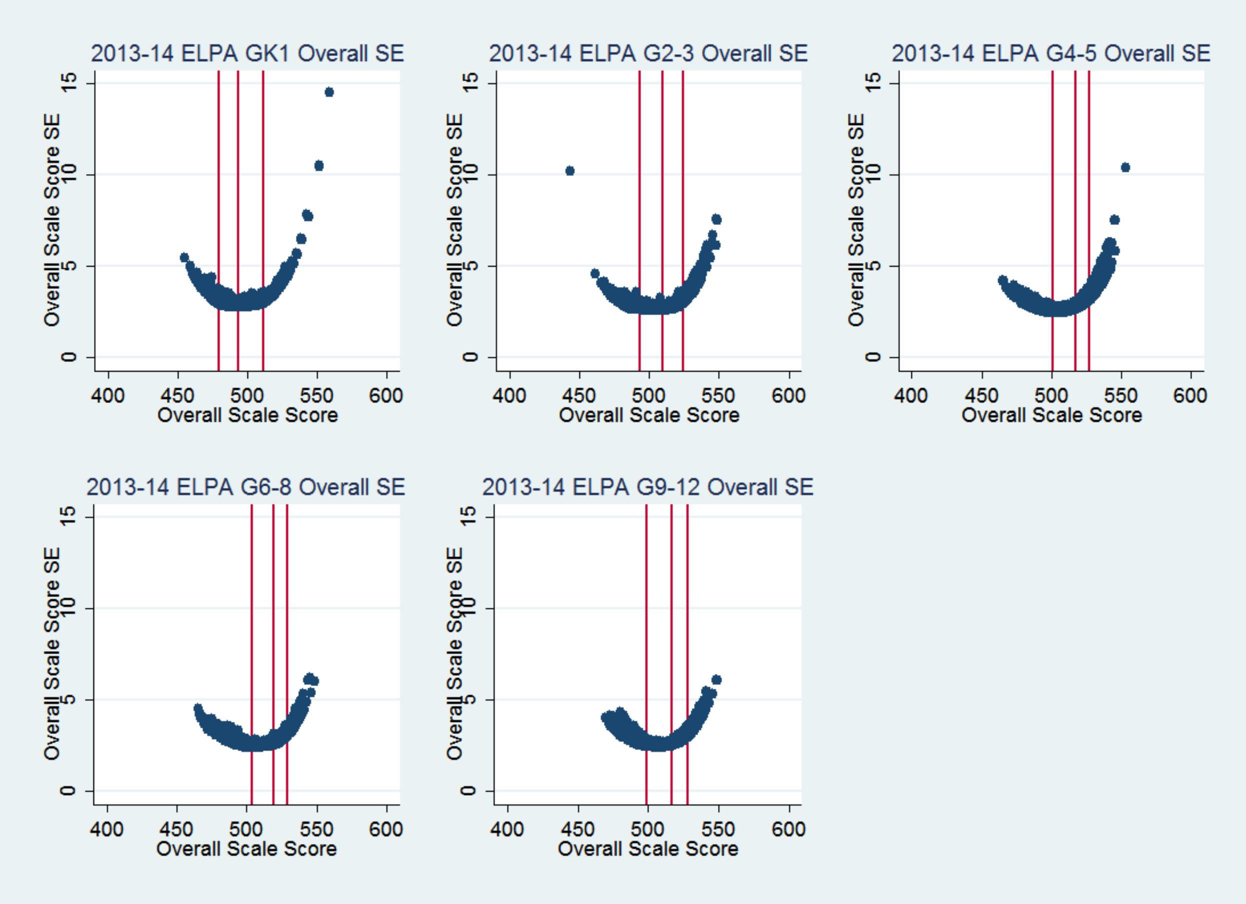
Table 13. Average Conditional Standard Error of Measurement at the threshold of Proficiency Levels, 2013–2014

Grade	Early Intermediate	Intermediate	Early Advanced	Advanced
K	3.07	2.87	2.85	2.90
1	2.9	2.86	3.07	3.81
2	2.74	2.67	2.69	2.89
3	2.69	2.7	3.06	3.41
4	2.65	2.53	2.63	2.86
5	2.73	2.57	2.78	3.31
6	2.66	2.57	2.51	2.68
7	2.87	2.53	2.65	2.96
8	3.35	2.54	2.71	3.11
9	2.83	2.57	2.48	2.69
10	2.73	2.61	2.54	2.94
11	3.00	2.54	2.51	2.89
12	2.7	2.59	2.53	2.88

7.1.2 Standard Error Curves

Figure 1 presents the conditional standard error of measurement across the range of ability by grade band for ELPA scores obtained from students who took the full length of the tests (not exempted from any domains) in 2013–2014. The item selection algorithm selected the items efficiently, matching to each student's ability while matching to the test blueprints, with the same precision across the range of abilities for all students. The standard error curves suggest that students are measured with a very high degree of precision although larger standard errors are observed at the higher ends of the score distribution than at the lower ends. There were 17 K-1 students, nine grades 2-3, seven grade 5 students, and three grade 7 students who had standard errors larger 6. This occurs because the item pools currently have a shortage of items that are better targeted toward these higher achieving individuals. The red line in the graph indicates the 10th percentile, the 50th percentile and the 90th percentile of the scale scores.

Figure 1. Conditional Standard Error of Measurement for Overall Scale Scores, 2013–2014



7.2 Reliability of Achievement Classification

When student performance is reported in terms of performance categories, a reliability index is computed in terms of the probabilities of consistent classification of students as specified in standard 2.15 in the *Standards for Educational and Psychological Testing* (AERA, 1999). This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications are estimated on a single-form test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model. (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995). For the computer-adaptive testing, because the adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability while meeting test blueprint requirements, the consistency classification is based on all sets of items administered across students.

The classification index can be examined for the decision accuracy and the decision consistency. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Decision consistency refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made on the basis of an alternate form (another set of adaptively administered items given the same ability), that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test forms.

In reality, the true ability is unknown, and students do not take an alternate, equivalent form; therefore, the classification accuracy and consistency is estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with a measurement error.

For the *i*th student, the student's estimated ability is $\hat{\theta}_i$ with a standard error, $se(\hat{\theta}_i)$, and the estimated ability is distributed, as $\hat{\theta}_i \sim N\left(\theta_i, se(\hat{\theta}_i)\right)$, assuming a normal distribution, where $\theta_i$is the unknown true ability of the *i*th student. The probability of the true score *at or above* the cut score is estimated as

$$p(\theta_i \geq \theta_c) = p\left(\frac{\theta_i - \hat{\theta}_i}{e(\hat{\theta}_i)} \geq \frac{\theta_c - \hat{\theta}_i}{e(\hat{\theta}_i)}\right) = p\left(\frac{\hat{\theta}_i - \theta_i}{e(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_c}{e(\hat{\theta}_i)}\right) = \Phi\left(\frac{\hat{\theta}_i - \theta_c}{e(\hat{\theta}_i)}\right).$$

Similarly, the probability of the true score being *below* the cutscore is estimated as

$$p(\theta_i < \theta_c) = 1 - \Phi\left(\frac{\hat{\theta}_i - \theta_c}{e(\hat{\theta}_i)}\right).$$

7.2.1 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se(\hat{\theta}_i)\right)$, we can estimate the above probabilities directly using the likelihood function. The likelihood function of theta given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated theta is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

The probability of a student being classified *at or above* the cut score, θ_c, given the student's item scores can be estimated as $P\left(\theta \geq \theta_c \mid r, \mathbf{b}\right) = \frac{\int_{\theta_c}^{+\infty} L(\theta \mid \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta \mid \mathbf{z}, \mathbf{b}) d\theta}$, where, the likelihood function is

$$L(\theta\,|\,\mathbf{z},\mathbf{b}) = \prod_{i\in MC}\left(\frac{Exp(z_i\theta - b_i z_i)}{1+Exp(\theta - b_i)}\right)\prod_{i\in CR}\left(\frac{Exp(z_i\theta - \sum_{k=1}^{z_i} b_k)}{1+\sum_{i=1}^{K_i} Exp(\sum_{k=1}^{i}(\theta - b_k))}\right)$$

$$\propto Exp(r\theta)\prod_{i\in MC}\left(\frac{1}{1+Exp(\theta - b_i)}\right)\prod_{i\in CR}\left(\frac{1}{1+\sum_{i=1}^{K_i} Exp(\sum_{k=1}^{i}(\theta - b_k))}\right).$$

Similarly, we can estimate the probability of *below* the cut score as:

$$P(\theta < \theta_c | r, \mathbf{b}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\mathbf{z},\mathbf{b})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta}.$$

In Figure 2, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassification (*false positive*) occurs when, for example, a student who actually achieved Early Advanced level on the true score, is classified incorrectly as achieving Advanced. N_{11} represents the expected numbers of students who are truly above the cut score, N_{01} represents the expected number of students falsely above the cut score, N_{00} represents the expected number of students truly below the cut score, and N_{10} represents the number of students falsely below the cut score.

Figure 2. Classification Accuracy

		Classification on a Form Actually Taken	
		Above the Cut score	**Below the Cut score**
Classification on True Score	**Above the** Cut score	N_{11} *(Truly above the cut)*	N_{10} (*False negative*)
	Below the Cut score	N_{01} (*False positive*)	N_{00} (*Truly below the cut*)

where

$$N_{11} = \sum\nolimits_{i\in N_1} P(\theta_i \geq \theta_c \,|\, r,\mathbf{b}),$$
$$N_{01} = \sum\nolimits_{i\in N_1} P(\theta_i < \theta_c \,|\, r,\mathbf{b}),$$
$$N_{00} = \sum\nolimits_{i\in N_0} P(\theta_i < \theta_c \,|\, r,\mathbf{b}), \text{ and}$$
$$N_{10} = \sum\nolimits_{i\in N_0} P(\theta_i \geq \theta_c \,|\, r,\mathbf{b}).$$

Where N_1 contains the students with estimated $\hat{\theta}_i$ being *at and above* the cut score, and N_0 contains the students with estimated $\hat{\theta}_i$ being *below* the cut score. The accuracy index is then computed as $\frac{N_{11}+N_{00}}{N}$, with $N = N_1 + N_0$

7.2.2 Classification consistency

Consistent classification occurs (Figure 3) when two forms agree on the classification of a student as either *at and above,* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Figure 3. Classification Consistency

		CLASSIFICATION ON THE 2ND FORM TAKEN	
		ABOVE THE CUT SCORE	BELOW THE CUT SCORE
Classification on the 1st Form Taken	**Above the Cut score**	*N*11 (Consistently above the cut)	*N*10 (Inconsistent)
	Below the Cut score	*N*01 (Inconsistent)	*N*00 (Consistently below the cut)

To estimate the consistency, we assume the students are tested twice independently hence, the probability of the student being classified as *at or above* the cut score θ_c in both tests can be estimated as $P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c) = P(\theta_1 \geq \theta_c)P(\theta_2 \geq \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})}\right)^2$.

Similarly, the probability of consistency and inconsistency can be estimated based on a student's item scores and the item parameters.

The probability of consistency for *at or above* the cut score is estimated as

$$P\left(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c \,|\, r, \mathbf{b}\right) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta}\right)^2$$

The probability of consistency for *below* the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c | r, \mathbf{b}) = \left(\frac{\int_{-\infty}^{\theta_c} L(\theta|\mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z}, \mathbf{b}) d\theta}\right)^2$$

The probability of inconsistency is estimated as

$$P\left(\theta_1 \geq \theta_c, \theta_2 < \theta_c \,|\, r, \mathbf{b}\right) = \frac{\int_{\theta_c}^{+\infty} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta \int_{-\infty}^{\theta_c} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta}{\left(\int_{-\infty}^{+\infty} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta\right)^2}, \text{ and}$$

$$P\left(\theta_1 < \theta_c, \theta_2 \geq \theta_c \,|\, r, \mathbf{b}\right) = \frac{\int_{\theta_c}^{+\infty} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta \int_{-\infty}^{\theta_c} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta}{\left(\int_{-\infty}^{+\infty} L(\theta \,|\, \mathbf{z}, \mathbf{b}) d\theta\right)^2}.$$

The consistent index is computed as $\frac{N_{11}+N_{00}}{N}$, where

$$N_{11}=\sum_{i=1}^{N}P\left(\theta_{i,1}\geq\theta_c,\theta_{i,2}\geq\theta_c\mid r,\mathbf{b}\right),$$

$$N_{10}=\sum_{i=1}^{N}P\left(\theta_{i,1}\geq\theta_c,\theta_{i,2}<\theta_c\mid r,\mathbf{b}\right),$$

$$N_{00}=\sum_{i=1}^{N}P\left(\theta_{i,1}<\theta_c,\theta_{i,2}<\theta_c\mid r,\mathbf{b}\right),$$

$$N_{01}=\sum_{i=1}^{N}P\left(\theta_{i,1}<\theta_c,\theta_{i,2}\geq\theta_c\mid r,\mathbf{b}\right),\text{ and}$$

$$N=N_{11}+N_{10}+N_{01}+N_{00}.$$

Table 14 presents the decision accuracy and consistency indexes 2013–2014. Accuracy classifications are slightly higher (1%–4%) than the consistency classifications in all performance standards. The consistency classification rate can be lower because the consistency is based on two tests with measurement errors while the accuracy is based on one test with a measurement error and the true score. The classification index ranged from 90% to 99% for the decision accuracy, and from 86% to 99% for the decision consistency across all grades. The accuracy and consistency indexes for each performance standard are higher for the standards with smaller standard error. The better the test is targeted to the student's ability, the higher the classification index is. For the Advanced standard, however, although the standard error is large, the accuracy and consistency rates are high, especially in grades K–3 because of the smaller number of tests classified in the Advanced level. If there are only a few students above a cut, most of the students are well below the cut with a high true negative rate (% below the Advanced standard), which produces a high consistency index.

Table 14. Decision Accuracy and Consistency Indexes for Performance Standards, 2013–2014

Grade	Early Intermediate	Intermediate	Early Advanced	Advanced
		Accuracy (%)		
K	96%	91%	94%	97%
1	91%	91%	94%	98%
2	96%	93%	94%	97%
3	97%	93%	92%	94%
4	99%	96%	91%	91%
5	99%	96%	92%	90%
6	99%	98%	91%	90%
7	99%	97%	91%	90%
8	99%	97%	91%	91%
9	98%	98%	93%	93%
10	99%	98%	94%	91%
11	99%	98%	94%	90%
12	99%	98%	93%	91%
		Consistency (%)		
K	94%	87%	92%	96%
1	87%	87%	92%	98%
2	95%	90%	91%	96%
3	96%	90%	89%	92%
4	98%	95%	88%	88%
5	98%	95%	89%	86%
6	99%	97%	88%	86%
7	99%	95%	88%	86%
8	99%	95%	87%	88%
9	98%	97%	90%	90%
10	98%	97%	91%	87%
11	99%	97%	92%	87%
12	99%	97%	91%	88%

7.3 Blueprint Match

Blueprints specify a range of items to be administered in each content domain and item type (multiple-choice items and constructed-response items) in each grade band. For the speaking domain, one to four difficult items are specified to make sure difficult items are selected in each speaking test. Table 15 presents the blueprint specifications in each grade band. All delivered tests satisfied the blueprint specifications 100%.

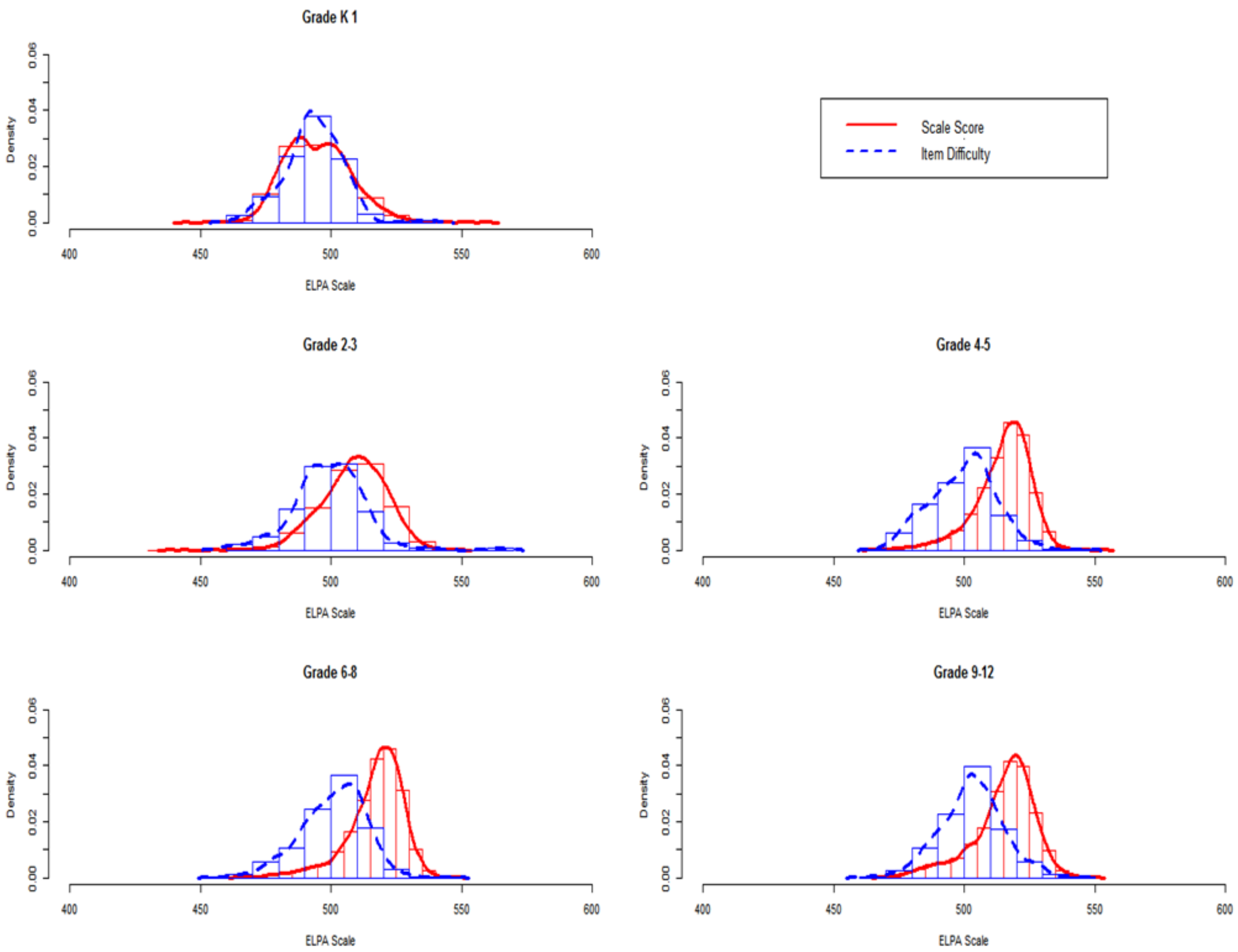
Table 15. ELPA Blueprint Specifications, 2013–2014

Domain/Affinity Group	Grade K–1		Grade 2–3		Grade 4–5		Grade 6–8		Grade 9–12	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Total Items	48	48	56	56	53	53	53	53	53	53
SPEAKING	8	8	6	6	3	3	3	3	3	3
Difficult Items	1	3	1	3	1	3	1	3	1	3
Not Difficult Items	0	7	0	5	0	2	0	2	0	2
LISTENING	10	15	10	15	14	20	14	20	14	20
READING	10	15	10	15	14	20	14	20	14	20
WRITING	14	19	23	32	12	14	12	14	12	14
WER	0	0	0	0	4	4	4	4	4	4

7.4 Student Ability–Item Difficulty Distribution for the 2013–2014 Operational Item Pool

Figure 4 displays the empirical distribution of the student scaled scores in the 2013–2014 administration and the distribution of the ELPA item difficulty parameters in the operational pool. The distributions in grades 2 and above indicate that the pool includes easier items than the ability of students in the tested population. The future field-test item development in all content domains is targeted to augment items in the areas that need improvement.

Figure 4. ELPA Student Ability–Item Difficulty Distribution, 2013–2014



Correlations Among Reporting Category Scores

ELPA scores are reported for the overall test performance and seven reporting categories. Table 16 presents the correlation coefficients among the total and reporting category scores. The pattern of correlations among the reading, speaking, writing and speaking is similar in all grade bands, ranges from 0.52 to 0.94. The correlation coefficients between grammar and illocution scores with other ELPA domains are smaller, ranges between 0.45 (for K-3) to .89. Among the lower grades, only two items (a total of four points) per test was used for estimating grammatical competence and illocutionary competence.

Table 16. Correlations Among Reporting Category Scores, 2013–2014

Content Strand	Overall	Listening	Reading	Speaking	Writing	Comprehension	Grammatical Competence	Illocutionary Competence
Grade K–1								
Overall	1.00	0.80	0.88	0.78	0.90	0.95	0.66	0.64
Listening		1.00	0.60	0.59	0.60	0.87	0.47	0.46
Reading			1.00	0.54	0.81	0.92	0.48	0.45
Speaking				1.00	0.56	0.63	0.82	0.81
Writing					1.00	0.80	0.49	0.46
Comprehension						1.00	0.53	0.51
Grammatical							1.00	0.81
Illocutionary								1.00
Grade 2–3								
Overall	1.00	0.77	0.90	0.77	0.94	0.93	0.60	0.67
Listening		1.00	0.62	0.58	0.60	0.85	0.45	0.50
Reading			1.00	0.60	0.80	0.94	0.44	0.53
Speaking				1.00	0.61	0.65	0.83	0.87
Writing					1.00	0.79	0.45	0.52
Comprehension						1.00	0.49	0.57
Grammatical							1.00	0.75
Illocutionary								1.00
Grade 4–5								
Overall	1.00	0.82	0.88	0.74	0.88	0.94	0.82	0.84
Listening		1.00	0.66	0.54	0.61	0.88	0.56	0.59
Reading			1.00	0.52	0.68	0.94	0.60	0.62
Speaking				1.00	0.58	0.58	0.78	0.78
Writing					1.00	0.71	0.84	0.85
Comprehension						1.00	0.64	0.67
Grammatical							1.00	0.79
Illocutionary								1.00
Grade 6–8								
Overall	1.00	0.85	0.86	0.75	0.89	0.94	0.84	0.86
Listening		1.00	0.67	0.52	0.67	0.90	0.61	0.62
Reading			1.00	0.51	0.69	0.92	0.60	0.62
Speaking				1.00	0.61	0.57	0.81	0.84
Writing					1.00	0.74	0.84	0.85
Comprehension						1.00	0.66	0.68
Grammatical							1.00	0.82
Illocutionary								1.00
Grade 9–12								
Overall	1.00	0.89	0.87	0.77	0.90	0.95	0.87	0.89
Listening		1.00	0.73	0.57	0.74	0.93	0.67	0.70
Reading			1.00	0.56	0.71	0.93	0.64	0.68
Speaking				1.00	0.61	0.61	0.85	0.86
Writing					1.00	0.78	0.83	0.84
Comprehension						1.00	0.71	0.74
Grammatical							1.00	0.85
Illocutionary								1.00

7.5 Online ELPA Scoring

A student's score for the adaptive ELPA depends on two factors: the number of items the student answers correctly and the difficulty of these items. In an adaptive assessment, each time a student answers an item, that item is scored, and the selection of subsequent items is based on how the student performed on earlier items. The first few items are selected to match to an average ELPA student because no previous score exists. As a student answers items correctly, the adaptive system assigns the student more difficult items. When a student answers items incorrectly, he or she will be given less difficult items. The online test delivery system administers the test adapting to each student's performance, while maintaining accurate representation of the required knowledge and skills in content breadth and depth specified in the test blueprints, and provides precise estimates of each student's true achievement level across the range of proficiency.

Test items are selected from the precalibrated item bank using a Rasch model to best match the ability level of each student. Student ability estimates are obtained by indexing items by *i*. The likelihood function based on the *j*th person's score pattern is

$$\mathrm{L_j}\left(\theta|z_j, b'_1, \dots b'_{k_j}\right) = \prod_{i=1}^{k_j} p_i(z_{ji}|\theta, b_{i,1}, \dots, b_{i,m_i}) \quad (3)$$

where $b'_1 = (b_{i,1}, \dots, b_{i,m_i})$ is the parameter vector of the *i*th item, m_i is the maximum possible score of this item, and the product is computed over only the k_j items presented to student *j*. Depending on the item type, the probability $p_i(z_{ji}|\theta, b_{i,1}, \dots, b_{i,m_i})$ takes either the form of a dichotomously scored item (in which case, we only have $b_{i,1}$, which can be simply written as b_i), or the form based on Masters' partial credit model for the polytomous items.

In case of dichotomously scored items, we have

$$p_i(z_{ji}|\theta, b_i) = \left\{ \begin{array}{c} \dfrac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} = p_i, if\ z_{ji} = 1 \\ \dfrac{1}{1 + \exp(\theta - b_i)} = 1 - p_i, if\ z_{ji} = 0 \end{array} \right\}$$

and in case of polytomous items,

$$p_i(z_{ji}|\theta, b_{i,1}, \dots, b_{i,m_i}) = \left\{ \begin{array}{c} \dfrac{\exp(\sum_{r=1}^{z_{ji}}(\theta - b_{i,r}))}{s_i(\theta, b_{i,1}, \dots, b_{i,m_i})}, if\ z_{ji} > 0 \\ \dfrac{1}{s_i(\theta, b_{i,1}, \dots, b_{i,m_i})}, if\ z_{ji} = 0 \end{array} \right\}$$

where $s_i\left(\theta, b_{i,1}, \dots, b_{i,m_i}\right) = 1 + \sum_{l=1}^{m_i} \exp(\sum_{r=1}^{l}(\theta - b_{i,r})$

and the log likelihood function is

$$l_i\left(\theta|z_j, b_{i,1}, \dots, b_{i,m_i}\right) = \log(L_j(\theta|z_j, b_{i,1}, \dots, b_{i,m_i})) = \sum_{i=1}^{k} \log(p_i(\ z_{ij}|\theta, b_{i,1}, \dots, b_{i,m_i}))\ . \ (4)$$

The ability θ is estimated by maximizing the log likelihood function defined in Equation (4), and the standard error of measurement (SEM) is approximated by the square root of the inverse of the Fisher information evaluated at the maximum likelihood estimate (MLE) of θ.

The student's ability (theta) estimates are linearly transformed to scaled scores. The scaled scores are mapped into five performance levels using four performance standards (i.e., cut scores).

7.5.1 Rules for Zero and Perfect Scores

In item response theory (IRT) maximum-likelihood ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. In such cases, MLEs are generated by adding ±0.5 to the zero or perfect raw scores, respectively, and maximizing conditional on the adjusted raw score.

7.6 Attemptedness Rule

Score reports are produced for the complete tests only. A completed test is defined as a test that was submitted (by clicking the submit button in the test) or a test where all the questions were answered but the test was not submitted by the student.

8. Summary of Student Performance

The 2013–2014 state summary results for the average scaled scores and the percentage of students in each performance level for overall performance and content domains are presented in Tables 17. Table 18 presents the distribution of the 2013–2014 tests if 2012–13 cut scores had been applied. A review of both tables shows that the largest change after the revision of the cut scores was found at the Advanced level of 9-12 students. With the 2013–2014 revised cut score, about 27 percent of high school students attained the Advanced level. With the 2012–13 cut scores, only about 12 percent of the same students would attain this performance level. Table 19 shows the percentage distribution by the subscales using 2013–2014 cut scores. The 2013–2014 scaled score distribution for each grade band is shown in Figure 5.

Table 17. Overall Mean Scaled Scores and Percentage of Students in Each Performance Level for the Total Test, Using 2013–14 Cut scores for Performance Levels

Grade	N	Mean	Std	% Beginning	% Early Intermediate	% Intermediate	% Early Advanced	% Advanced
KG	8021	487	8.57	21	45	20	12	3
1	8085	503	10.93	14	36	31	15	5
2	7573	505	10.99	13	30	36	15	6
3	7069	514	11.64	12	24	36	15	14
4	6469	514	10.25	5	10	28	36	22
5	4621	517	10.62	4	12	17	41	26
6	3117	517	10.25	3	6	28	29	34
7	2256	517	11.01	5	11	27	28	29
8	1607	518	11.45	6	10	32	30	22
9	1297	512	12.07	10	5	27	41	17
10	1320	515	11.68	6	5	21	42	26
11	1174	517	11.34	4	3	19	38	36
12	1237	516	11.49	6	3	23	38	30

Table 18. Overall Mean Scaled Scores and Percentage of Students in Each Performance Level for the Total Test, Using 2012–13 Cut scores for Performance Levels

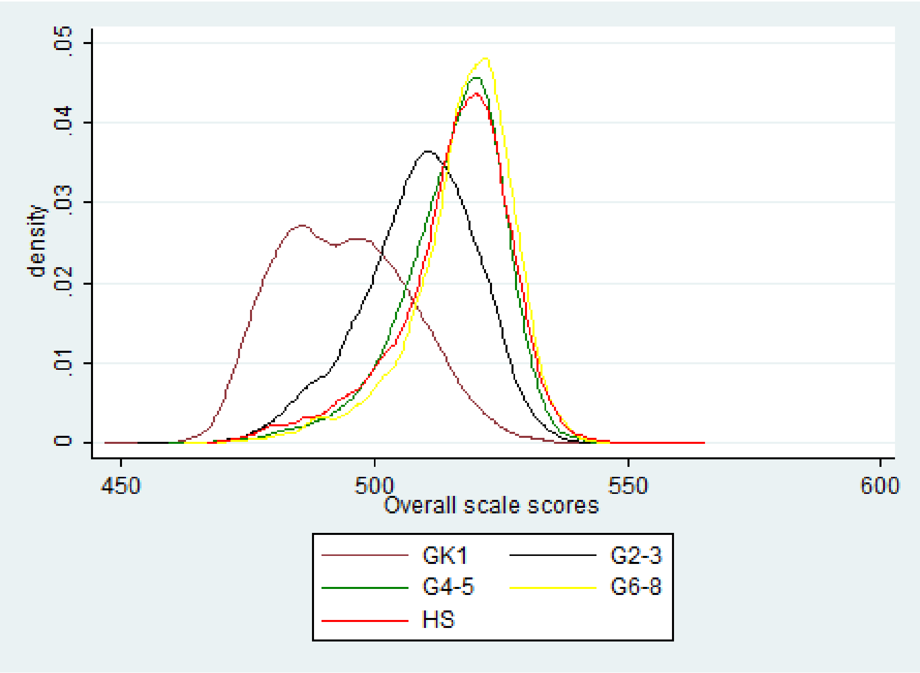
Grade	N	Mean	Std	% Beginning	% Early Intermediate	% Intermediate	% Early Advanced	% Advanced
KG	8021	487	8.57	25	45	18	10	2
1	8085	503	10.93	15	51	19	11	4
2	7573	505	10.99	18	39	21	18	4
3	7069	514	11.64	13	33	25	21	8
4	6469	514	10.25	7	16	20	31	27
5	4621	517	10.62	5	11	21	33	31
6	3117	517	10.25	5	7	22	32	34
7	2256	517	11.01	6	9	25	32	29
8	1607	518	11.45	7	8	24	35	27
9	1297	512	12.07	8	8	32	43	9
10	1320	515	11.68	6	6	30	45	13
11	1174	517	11.34	4	5	24	51	16
12	1237	516	11.49	7	6	27	50	9

Table 19. Mean Scaled Scores and Percentage of Students in Each Performance Level for Reporting Categories, by Grade Bands 2013–2014

Grade	N	Mean Scaled Score	SD	% Beginning	% Early Intermediate	% Intermediate	% Early Advanced	% Advanced
Grade K–1								
Listening	16100	497	14.66	18	30	22	17	13
Reading	16087	494	14.51	28	33	20	14	5
Writing	16088	495	15.22	24	37	19	12	8
Speaking	16052	495	16.95	26	26	23	15	10
Comprehension	16106	495	12.85	19	36	27	14	4
Grammatical	16033	498	15.2	27	26	15	16	16
Illocutionary	16033	499	15.57	23	23	18	20	17
Grade 2–3								
Listening	14634	510	12.48	11	27	35	15	13
Reading	14598	509	15.2	17	26	29	13	15
Writing	14601	509	14.02	15	24	30	16	15
Speaking	14599	509	16.42	18	26	26	13	18
Comprehension	14642	509	12.77	11	29	36	12	11
Grammatical	14575	507	12.83	20	26	39	11	3
Illocutionary	14575	510	15.95	22	26	17	10	25
Grade 4–5								
Listening	11088	515	11.51	4	12	27	33	23
Reading	11012	514	11.09	6	14	26	33	21
Writing	11016	518	15.14	7	10	19	26	38
Speaking	11051	518	16.75	7	12	18	23	40
Comprehension	11090	514	10.14	4	12	28	36	19
Grammatical	11083	521	18.08	7	10	15	21	47
Illocutionary	11083	516	14.13	7	10	22	28	32
Grade 6–8								
Listening	6980	516	11.02	4	9	34	26	26
Reading	6972	515	12.15	5	14	34	21	25

Writing	6913	521	15.05	5	8	20	19	47
Speaking	6923	518	16.6	10	10	19	17	44
Comprehension	6948	516	10.38	4	11	36	27	22
Grammatical	6976	524	17.5	6	8	16	12	58
Illocutionary	6972	518	14.64	8	9	24	21	38
Grades 9-12									
Listening	5024	514	12.64	7	5	28	37	23
Reading	5020	515	11.74	5	5	28	37	26
Writing	5018	517	15.53	8	4	20	29	39
Speaking	5015	518	18.57	13	4	17	22	44
Comprehension	5028	514	11.25	5	5	28	39	23
Grammatical	5024	519	18.34	9	4	18	20	50
Illocutionary	5024	516	15.86	10	3	19	30	37

Figure 5. ELPA Scaled Score Distribution, 2013–2014



REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and psychological testing.* Washington, DC.

Livingston, S.A., & Lewis, C. (1995). Estimating the consistency and accuracy of classificatons based on test scores. *Journal of Educational Measurement*, *32*, 179-197.

Livingston, S.A., & Wingersky, M.S. (1979). *Assessing the Reliability of Tests Used to Make Pass/Fail Decisions.* ETS Policy and Research Reports.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: LEA.

Kish, L. (1965). *Survey sampling.* New York: John Wiley and Sons.

Linacre, J. M. (2011). *WINSTEPS Rasch-Model computer program*. Chicago: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Subkoviak, M. (1976). Estimating Reliability from a Single Administration of a Criterion-Referenced Test. *Journal of Educational Measurement*, *13*, 265-276.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. (1979). *Best test design. Rasch measurement.* Chicago: MESA Press.