Alt ELPA
Alternate English Language Proficiency Assessment

Alt ELPA:

Embedded Standard Setting (ESS) and Cut Score Development

Executive Summary

Table of Contents

About the Alt ELPA

The Alt ELPA is a new summative assessment designed to measure the English language proficiency of English Learners with the most significant cognitive disabilities. The Alt ELPA's purpose is to provide fair and valid information on the English language proficiency of this historically underserved group of students, in compliance with the Every Student Succeeds Act of 2015. All students with a significant cognitive disability who are identified as English Learners are eligible to, and required to, participate in annual Alt ELPA testing. The Alt ELPA is administered annually to English Learners with the most significant cognitive disabilities in grades K-12.

The Alt ELPA has been thoughtfully designed and developed by a collaborative group of ten state departments of education[1], led by the Iowa Department of Education, and national assessment experts. The Iowa Department of Education was awarded ~$7.9M in 2019, under the U.S. Department of Education's Competitive Grants for State Assessments Program, to develop and launch this assessment. The assessment was designed in partnership with the states by the Center for Research, Evaluation, Standards, and Student Testing (CRESST) at UCLA. Creative Measurement Solutions, Inc. (CMS), a leading educational measurement consultancy, partnered with CRESST to lead the participating states through a design and development process that included Embedded Standard Setting, a methodology that transforms the traditional "stand-alone" standard setting process into a continuous, active component of designing an assessment.

The assessment consists of test items that measure the two modalities of English language development: the Receptive modality, which measures listening and reading, and the Productive modality, which measures speaking and writing. We categorize English language proficiency on the Alt ELPA as a combination of students' performance in both the Receptive and Productive modalities.

The Alt ELPA has 3 overall proficiency determination categories, as described below:

- Proficient – Students show a level of English language proficiency reflected in the Alternate ELP standards that enables full participation or only slightly limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards. This is indicated on the Alt ELPA by attaining Level 3 or higher in all modalities. Once Proficient on the Alt ELPA, students may be considered for reclassification.

[1] The following states participated in the development of the Alt ELPA: Arizona, Arkansas, Connecticut, Iowa, Louisiana, Nebraska, New York, Ohio, Oregon, and West Virginia.

- Progressing – Students show a level of English language proficiency reflected in the Alternate ELP standards that moderately limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards. This is indicated on the Alt ELPA by attaining above Level 1 and below Level 3 in at least one modality. Students scoring Progressing on the Alt ELPA are eligible for ongoing program support.
- Emerging – Students show a level of English language proficiency reflected in the Alternate ELP standards that significantly limits participation in the grade-appropriate classroom activities reflected in the Alternate Academic standards. This is indicated on the Alt ELPA by attaining Level 1 in all modalities. Students scoring Emerging on the Alt ELPA are eligible for ongoing program support.

Students are placed into the proficiency categories above based on their scores in the Receptive and Productive modality. Performance in a modality is described by four performance levels:

- Level 1 – Beginning
- Level 2 – Intermediate
- Level 3 – Early Advanced
- Level 4 – Advanced

A modality performance level of 3 or 4 indicates that the student is demonstrating that they have the English language skills in that modality, as described in the Alternate English language proficiency standards, to participate in grade-appropriate academic content, as measures by the state's alternate content standards. Students who achieve Level 3 or 4 in both modalities are considered Proficient, and are eligible to be exited from English language services.

The receptive and productive modalities are divided into the four performance levels listed above by cut scores, which are selected points on the score scale of a test or sub-test. Cut scores divide the continuum of student performance into performance levels. The Embedded Standard Setting process is one of the methodologies used to identify the assessment's cut scores.

Methodology & Processes

Establishing the appropriate Alt ELPA cut scores for in modalities and grade levels is a multi-step, iterative process. This process reflects the Embedded Standard Setting approach, and includes item specification, item writing, item review, field testing, item bank calibration, estimation of cut scores, application of expert judgement to cut scores (also called "smoothing"), Inconsistent Item Review, and state approval of cut scores. These processes all

work on concert to establish cut scores that reflect the performance expectations in the English Language Proficiency standards, and to ensure that the Alt ELPA measures the intended construct and provides valid, meaningful assessment scores.

As an initial step, performance level descriptors are written by experts familiar with the instruction this group of students, states' alternate academic content standards, and the Alternate English language proficiency standards. Those performance level descriptors describe the language skills the student knows and can demonstrate on the assessment. Using those performance level descriptors, professional item writers construct assessment items, and each item is written to measure a specific "target", or performance level, of a standard.

Once the items are written, a panel of expert educators is convened to review and confirm that the item writers have met the target of measurement for each assessment item. Items that are judged to have met their targets are then field tested. **Field testing** is yet another step in validating the assessment and the accuracy of the measurement targets reflected in each item.

Using the empirical evidence – student testing data -- available from the field test, CRESST generated item statistics, calibrated the item bank and established **cut scores** for each modality. Using item response theory, the probability of test-takers at various levels of performance providing correct responses of an item can be estimated. Creative Measurement Solutions' analytical method considers the proficiency level descriptors and the judgements of subject matter experts in addition to item response data when calculating the optimal cut scores. These analytical cut score estimates are not perfect, but their uncertainty can be quantified by calculating standard error values.

When the assessment is field tested, some items perform as expected, (e.g., they measure the performance level they are intended to measure). For some items, however, there are be discrepancies between the analytical results and the subject matter experts' judgements. It is at this point when a panel of expert educators convenes to reconcile these discrepancies. This is called **Inconsistent Item Review**, and the panel event takes place on August 2-3, 2023. The feedback and expert judgements of this panel of expert educators is used to further refine the assessment's items.

For example, the item writers may have indicated that an item is most appropriate for Level 2 learners, but field test data could show that even Level 1 learners provided correct responses with ease. This is considered an inconsistent item. The educators at the Inconsistent Item Review Workshop will review both the item writers' target of measurement and the assessment data, and will render a judgement of which performance level the item is actually targeting. This expert judgement is an important step in the process of establishing cut scores and preparing the assessment to be administered operationally.

Another element in the cut score setting process is the **Contrasting Groups Study**. In this study, teachers who regularly work with English Learners with significant cognitive disabilities are asked to rate their students' proficiency levels overall, and on each modality. This study was conducted at the same time the students took the field test. Ideally, the results from these surveys should be consistent with the empirical results from the test – the assessment data showing how students performed. For example, if a student scores "Proficient" on the Alt ELPA, their teacher should ideally have the same opinion about that student's language proficiency and their ability to participate in grade-level instruction without English Learner services.[2]

Results from the Alt ELPA Contrasting Groups Study were generally consistent with the empirical results. Teachers tended to rate productive skills lower than what students actually achieved on the test, although a representative from the Iowa Department of Education mentioned that this outcome is typical in Contrasting Groups Studies.

The Inconsistent Item Review Workshop and the Contrasting Groups study provide valuable insights into the accuracy of the cut scores. And analytically optimal cut scores are not always optimal in a practical sense. Additional evidence contributed by state education agency (SEA) stakeholders is also considered. SEAs must balance the precision and rigor of an assessment's cut scores with the impact of those cut scores on student outcomes. Adjustments to cut scores, commonly referred to as "smoothing", may need to be made if the proposed cut scores result in unusually high or low proportions of learners being classified into a particular category.

In the case of the Alt ELPA, smoothing of the cut scores allow states to administer an assessment that balances fairness and validity. With the original, analytically derived cut scores, the proportion of Alt ELPA test-takers in grades K and 1 demonstrating proficiency in productive skills was near zero. This made it effectively impossible for an English learner with the most significant cognitive disabilities to receive an overall proficient rating and thereby test out of EL services. And while the meaning and purpose of a test that is essentially impossible to pass can be called into question, test designers and SEAs have to make sure the Alt ELPA maintains its integrity by ensuring that students who are determined to be proficient do indeed have the language skills to engage in instruction alongside their peers, who are students with significant cognitive disabilities who are *not* English learners.

Thus, the states engaged in further discussion with CRESST and CMS about smoothing (e.g., adjusting) the assessment's cut scores. These small adjustments were done by means of **vertical articulation**, which is a method of adjusting the cut scores to account for incremental

[2] If this ideal case is realized, we have additional validity evidence for the Alt ELPA cut scores. If not, we have additional pieces of information for overview and inquiry regarding the Alt ELPA cut scores.

growth from one grade level to the next by expanding the cut scores to include the statistical margin of error. In response to this issue, states voted to adopt an adjustment to the cut scores that would increase the percentage of proficient grade K and 1 English learners from about 1% to about 3-4%. This **smoothing** of the cut scores also helps to align the assessment outcomes with the intended purpose and guiding principles of the assessment.

Additional evidence in support of this smoothing is that a 3-4% proficiency rate for grade K and 1 students is similar to that for the general ELPA population (students who are English learners but are not significantly cognitively disabled). CMS, CRESST, and SEAs also voiced that slightly conservative cut scores are beneficial for students in low grades so that students who are near, but not yet proficient, can continue receiving important English learner services and support for at least another year, allowing these students to advance their language skills and improving their likelihood of success with grade-appropriate content classrooms.

Alt ELPA Cut Scores

After the Alt ELPA state representatives approved the smoothing of the cut scores, CRESST transformed the IRT scores into scores on the Alt ELPA score reporting scale, which reports modality and domain scores on a scale of 0-99. This scale was chosen for its familiarity to students, families, and educators. Using a familiar scale increases the meaning and accessibility of Alt ELPA scores to the numerous stakeholders who will use these scores.

The Alt ELPA measures the four domains of language—listening, reading, speaking, and writing—as required in Federal law. The Alt ELPA reports student scores based on those four domains and two modalities of language: reception and production.

Reception, also called the "receptive modality", expresses how a student receives communication: via listening and reading. Production, also known as the "productive modality", describes how a student produces communication, via speaking and writing. The cut scores for each domain (listening, reading, speaking, and writing) are the same as the cut scores for the modality that domain is in (receptive or productive).

Cut scores for reporting are provided as a range of the lowest (minimum) and highest (maximum) cut scores for each performance level. The proposed cut scores for the Alt ELPA are listed in *Table 1 and Table 2.*

Table 1. Alt ELPA Receptive Modality Cut Scores for Grades K-12.

Grade	Receptive Modality (Listening and Reading)							
	L1		L2		L3		L4	
	Min	Max	Min	Max	Min	Max	Min	Max
KG	0	61	62	70	71	82	83	99
1	0	56	57	64	65	83	84	99
2	0	50	51	60	61	79	80	99
3	0	55	56	65	66	82	83	99
4	0	38	39	51	52	83	84	99
5	0	44	45	57	58	86	87	99
6	0	33	34	42	43	79	80	99
7	0	34	35	44	45	80	81	99
8	0	35	36	45	46	80	81	99
9	0	35	36	46	47	82	83	99
10	0	35	36	46	47	82	83	99
11	0	35	36	46	47	82	83	99
12	0	35	36	46	47	82	83	99

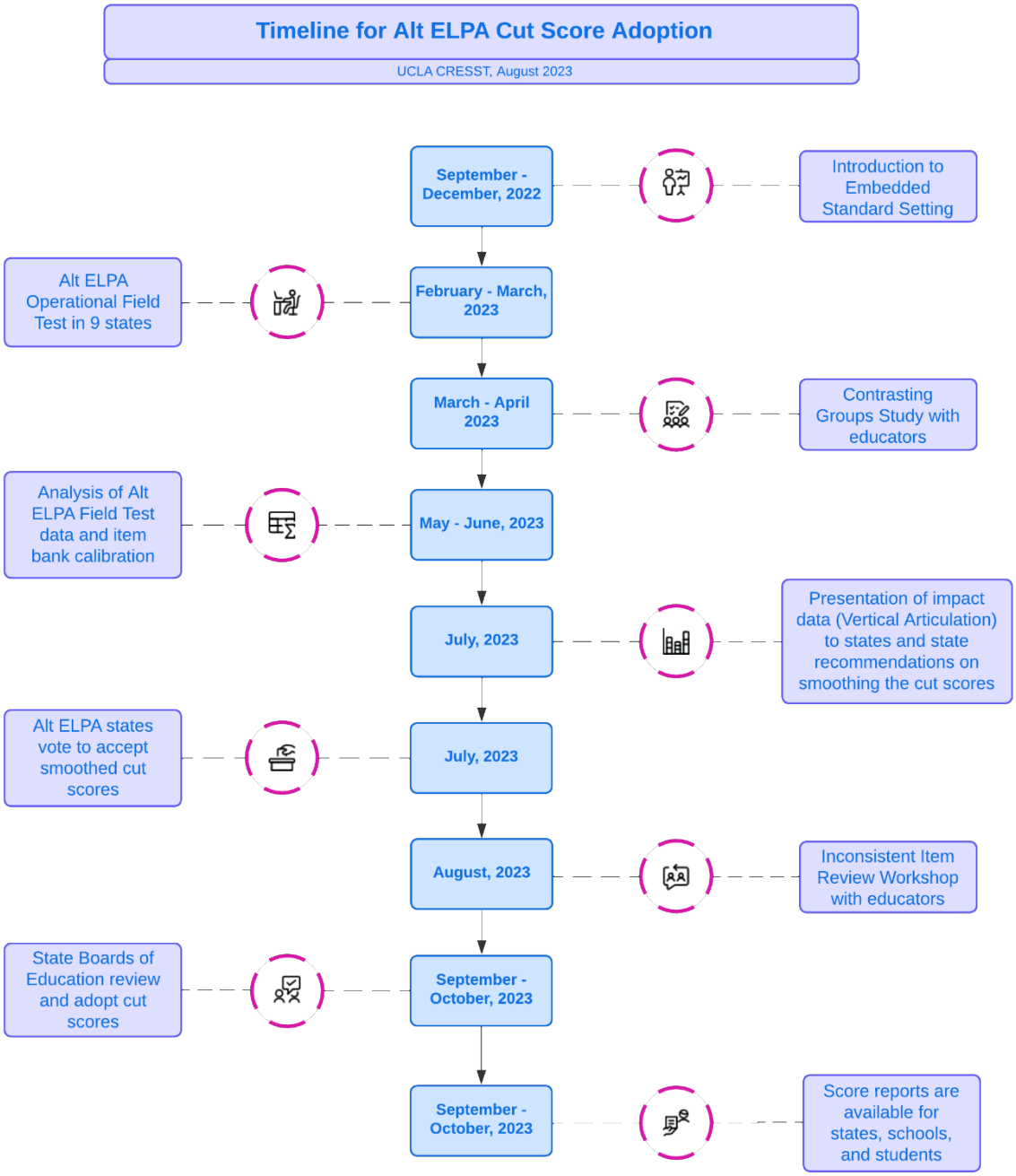
Table 2. Alt ELPA Productive Modality Cut Scores for Grades K-12

Grade	Productive Modality (Speaking and Writing)							
	L1		L2		L3		L4	
	Min	Max	Min	Max	Min	Max	Min	Max
KG	0	73	74	83	84	91	92	99
1	0	67	68	82	83	94	95	99
2	0	61	62	80	81	87	88	99
3	0	66	67	83	84	89	90	99
4	0	41	42	72	73	80	81	99
5	0	48	49	77	78	84	85	99
6	0	41	42	64	65	83	84	99
7	0	42	43	66	67	84	85	99
8	0	44	45	67	68	85	86	99
9	0	49	50	67	68	76	77	99
10	0	49	50	67	68	76	77	99
11	0	49	50	67	68	76	77	99
12	0	49	50	67	68	76	77	99

Next Steps for States

States planning to operationalize the Alt ELPA will review the assessment's cut scores with their state Boards of Education in September and October of 2023.

A timeline for the cut score adoption process, and a list of additional resources, are on the following pages.


Timeline for Alt ELPA Cut Score Adoption
UCLA CRESST, August 2023
September - December, 2022
Introduction to Embedded Standard Setting
Alt ELPA Operational Field Test in 9 states
February - March, 2023
March - April 2023
Contrasting Groups Study with educators
Analysis of Alt ELPA Field Test data and item bank calibration
May - June, 2023
July, 2023
Presentation of impact data (Vertical Articulation) to states and state recommendations on smoothing the cut scores
Alt ELPA states vote to accept smoothed cut scores
July, 2023
August, 2023
Inconsistent Item Review Workshop with educators
State Boards of Education review and adopt cut scores
September - October, 2023
September - October, 2023
Score reports are available for states, schools, and students

RESOURCES

- Alt ELPA Standard Setting Design document (to be finalized Fall 2023): Link
- Overall proficiency categories and modality reporting PLDs: Link