

## **The English Language Development Assessment (ELDA)**

Authors: Julia Lara (Consultant); Steve Ferrara, Mathina Calliope (American Institute for Research); Diana Sewell (Louisiana Department of Education); Phoebe Winter (Consultant); Rebecca Kopriva (C-SAVE); Michael Bunch, Kevin Joldersma, (Measurement Incorporated.)

### **Introduction to the Consortium**

The *No Child Left Behind Act of 2001* (NCLB; 2002) requires all states to assess the English proficiency of English language learners each school year. Under Title I and Title III of NCLB, states are required to measure the annual growth of students' English language development in reading, listening, writing, and speaking and in comprehension toward attainment of full English proficiency. Language Development Assessment (ELDA) was designed assess the development of proficiency in relation to English language proficiency standards of participating states.

To assist states in meeting these requirements, the Council of Chief State School Officers (CCSSO), along with states in the State Collaborative on Assessment and Student Standards for Limited English Proficient students (LEP-SCASS), solicited proposals from test development organizations.<sup>1</sup>

The LEP-SCASS and CCSSO selected the American Institutes for Research (AIR) to work collaboratively to develop English language proficiency assessments. The

---

<sup>1</sup> The CCSSO SCASS projects are networks of state education agency staff that combine their resources for the purpose of development assessment related tools and products that benefit the member states. There are 13 such state networks coordinated by CCSSO. Member states pay a yearly fee to the Council to defray the cost of travel, overnight accommodations, consultants, and administration. The LEP SCASS consortium was formed at the request of state education agency officials interested in developing procedures, products and services focused on ELL students. Since its inception the LEP SCASS staff and consultants have produced guides for scoring ELL student response to math and science items, a handbook for assessing ELL students, research papers on ELL assessment issues.

LEP-SCASS received an Enhanced Assessment Grant under Title VI (Section 6112) of No Child Left Behind (P.L. 107-110; NCLB) from the US Department of Education to fund development, validation, and implementation of an English proficiency assessment. During fall 2002 through December 2005, the LEP-SCASS, CCSSO, AIR, and Measurement Incorporated (MI) worked together to develop the English Language Development Assessment (ELDA). Nevada, as the lead state in the grant, and CCSSO managed the ELDA project for the LEP-SCASS. The project included outside consultants to evaluate the development process and provide design and technical advice, the Center for the Study of Assessment Validity and Evaluation (CSAVE) at the University of Maryland, who conducted validation studies, and Measurement Inc. responsible for K-2 development, administration, scoring and reporting of the tests. At the start of the ELDA development project, 18 states were members of the LEP SCASS. Thirteen states participated in the process of developing, field testing, validating, and implementing ELDA as an operational assessment.

The consortia determined that a valid English language proficiency assessment for K-2 English learners should rely upon observational data of English learners in natural classroom settings. For this reason, a separate test blueprint was developed for the K-2 ELDA forms and the 3-12 ELDA forms. It should be noted that the K-2 and the 3-12 versions of the ELDA are both driven by theories of academic language and are both aligned to participating states' ELP standards. The test development process for the ELDA 3-12 and the ELDA K-2 are described in separate sections below.

## **ELDA Grades 3-12**

### **The Theoretical Basis of the ELDA**

ELDA has been designed to assess the construct of “academic English” (Butler et al, 2004). The driving force—and departure of this assessment from many English language proficiency assessments—is NCLB’s requirement that students classified as English-language learners be assessed annually in their progress towards proficiency in academic English. For purposes of test design and development, we defined academic English as falling into one of two categories: (1) language used to convey curriculum-based, academic content, and (2) the language of the social environment of a school. The concept of academic English is evolving, and it is important to make the point that although the ELDA items and prompts are written in the language of the classroom and of the academic subjects listed below, items do not require skills in or knowledge of content in those subjects. The *concepts* are not being assessed; the students’ understanding of spoken and written texts *about* the concepts and ability to write and speak *about* the concepts are being assessed. Any content a student is expected to use is provided in the stimuli.

Three academic content topic areas constitute the foundation for selection of, and creation of the context for, ELDA test items in all four skills domains (Listening, Speaking, Reading, and Writing) of the test:

- English Language Arts
- Math, Science, and Technology
- Social Studies
- School environment

This assessment is informed by second-language development theory of communicative competence which posits that ELP tests should measure communicative and participatory language in the context of the classroom and that they should be age/grade appropriate. This test is a departure from existing ELP tests in that ELDA measures English language mastery along the language development continuum and for each linguistic domain. In addition, it attempts to measure mastery of “academic English.” Previous ELP assessments were designed to assist local educators with student placement decisions and measured low levels skills. Consequently, students were exited to the English only classroom before mastery of the English language skills. Once in the classroom, ELL students were unable to meet the linguistic demands of the classroom and were often labeled as poor performers. Thus, limited proficiency in English was confounded with knowledge of subject matter taught in the classroom. Moreover, these tests were not able to provide instructionally relevant information, nor were they aligned to state English language development standards.

### **Standards Used as a Base for Test Development**

The starting point for the ELDA test design was a synthesis of all state-level ELP standards that existed among the project’s participating states. Of 18 states that initially formed the LEP-SCASS membership, one-third had existing state ELP standards in each of the four domains of Listening, Speaking, Reading, and Writing.

The initial state ELP standards were carefully reviewed and merged by AIR staff. Then, project Steering Committee members agreed on a common core of standards for each domain by discussing standards they considered important and appropriate for ELLs in all LEP-SCASS states. They also considered which standards were appropriate at each

grade cluster. Some states used the ELDA ELP standards to guide the development, revision, analysis, and adoption of their own ELP standards. Other states used them to review their existing ELP standards and ensure alignment with ELDA ELP standards.

State academic content and achievement standards are mandated by the U.S. Department of Education under NCLB for three content areas: Reading/Language Arts, Mathematics, and Science. With reference to testing of ELL students under Title III of NCLB, the law requires that *English language proficiency standards* be aligned with challenging state academic content standards and student academic achievement standards as described in Title I.

In order to align test to standards, the CCSSO and AIR led a detailed and stakeholder-approved process of identifying ELP standards that could be used in test design, with standards yielding benchmarks and standards and benchmarks aligned with test items. In the case of academic content standards, the relationship to the test is less direct. Alignment between content standards and *an ELP assessment* is not implied in the non-regulatory guidance put forth by ED.<sup>2</sup>

### **Test Blueprint and Item Development**<sup>3</sup>

AIR test developers and psychometricians drafted test blueprints and item specifications for each domain and grade cluster. To develop items that measure ELDA's ELP standards as specified by the content specifications, AIR brought together a pool of item writers which included external item writers, NAEP foreign language item writers,

---

<sup>2</sup> Final Non Regulatory Guidance on the Title III State Formula Grant Program Standards, Assessment, Accountability. Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students, February, 2003.

<sup>3</sup> See Standards and specification document at: [www.ccsso.org/projects/elda/Research\\_Studies](http://www.ccsso.org/projects/elda/Research_Studies).

AIR content experts, and teachers who had experience with test development and were recommended by the LEP-SCASS states.

### *Listening*

The listening test for each of the three grade clusters (3-5, 6-8, and 9-12) is designed to be administered through a cassette tape or CD medium. All test items are four-option multiple choice. Listening texts impart information drawn from the four topic areas: English/Language Arts; Mathematics, Science, and Technology; Social Studies; and School-Environmental. Text topics within the academic domains were selected to avoid those that would typically be found on a grade-appropriate curriculum to ensure the assessment would measure comprehension, and not prior content-area knowledge. The texts are written, however, to reflect the discourse features typical of the domain. Grade clusters 3-5 and 6-8 operational forms contain a total of 50 test items each, grade cluster 9-12 a total of 60 multiple-choice items. High test item totals for operational forms are a product of a five-level scale of performance.

### *Speaking*

The Speaking assessment is designed to be administered through a cassette tape or CD medium, thus eliminating written discourse from the measurement of an oral-based construct. It also can be administered orally to individual students. A test booklet containing graphics provides the student with some visual contextualization of the prompts on cassette/CD. The graphics are designed to help the student structure a response. Student responses to the prompts are captured on an individual student cassette player for off-site scoring. In operational administrations of the speaking test, schools may opt for oral administrations with local scoring, although the test content remains the

same. Responses are scored on a 0-2 rubric. This rubric identifies rhetorical features (i.e., organization of ideas and information, use of discourse markers to support organization), appropriateness (i.e., relevance and completeness), quality and quantity of the response (i.e., development and specificity, adequacy of the response in addressing the task), and correctness (i.e., appropriate vocabulary and comprehensible pronunciation) of spoken responses. The rubric and benchmark responses also account for consideration of audience.

### *Reading*

The reading test for each of the three completed grade clusters has many of the design features of the listening test: four-option multiple choice test format, identical operational test item numbers in each cluster, identical approach to topic content selection of reading texts and similar distributions across the four academic and non-academic topic areas. The reading test is composed of three sections: Early Reading; Reading Instructions; and Reading Narrative, Descriptive, Expository, and Persuasive Texts. As it is for Listening, the Reading grade clusters 3-5 and 6-8 operational forms contain a total of 50 test items each and grade cluster 9-12 a total of 60 multiple-choice items. High test item totals for operational forms are a product of a five-level scale of performance.

### *Writing*

The writing test for grade clusters 3-5, 6-8, and 9-12 shares some features with the listening and reading components—distribution across topic areas and an emphasis on the language of the classroom—with a key difference: it also, logically, contains constructed response items. Three main sections comprise the writing test: multiple choice editing

and revising items (6-9 items per form), multiple choice planning and organizing items (6 items per form) and a combination of short and extended constructed response essay prompts (4-5 per form).

The editing and revising items are built around short stimuli, designed to simulate student writing. In some of these multiple choice items, relevant portions of these texts (a word or phrase), which may be grammatically incorrect, are underlined; students are asked to choose from options to replace the underlined text or to indicate that it already is correct. In other items, students are asked to choose an appropriate topic or concluding sentence or to provide missing information. The revising and editing items are designed to test students' ability to identify and correct sentence-level as well as text-level problems.

### **Pilot and Field Testing of Assessment Items and Tasks**

In May 2003, 31 schools in 12 states participated in a pilot test of the ELDA. The purpose of the pilot test was to determine whether test administration directions were clear for teachers and students, test administration procedures were feasible and efficient, and English language learners would be able to respond reasonably to the various item types. The pilot test included the reading, listening, writing, and speaking assessments for the three grade clusters. Schools were identified and recruited for participation so that the sample of schools was diverse in terms of type, size, location, and student demographics. Each school provided 10 students: five English language learners with low-mid English proficiency and five additional students that included a mix of English language learners with mid-high English proficiency, former limited English proficient students, and native English speakers. Participating students were drawn from each of grades 3-12 and reflect

a diverse mix of race/ethnicity, sex, native language, country of birth, time in the US, and time learning English.

Pilot test students came from more than 20 language backgrounds and more than 30 countries. Results from item analyses, student focus group reports, and teacher reports indicated that students understood test administration procedures and were able to give their best performances in all four skill domains. Test score reliabilities ranged from .77 to .92, similar to score reliabilities achieved in state content area assessments. Based on input from pilot test teachers, AIR and SCASS members revised administration procedures to make administering the test easier for teachers and taking the test easier for students. Analysis of results informed further item development.

A multi-state field test was conducted in spring 2004. The purposes of the 2004 field test were to (a) gather adequate data (i.e., 1,000 responses per item) to evaluate items and create ELDA score scales, (b) assemble operational form 1 of the Listening, Reading, Writing, and Speaking assessments for use in 2005, and (c) conduct special studies relevant to the validity of interpretations about English proficiency from ELDA test scores.

Both field test and operational administrations were conducted in spring 2005. A field test was conducted in five states: Georgia, Indiana, Kentucky, New Jersey, and Oklahoma. The primary purpose of the 2005 field tests was to yield data on items to assemble operational forms 2 and 3 of assessments for operational use beyond 2005. Six states administered operational form 1 of ELDA for operational purposes and reported results to meet No Child Left Behind requirements: Iowa, Louisiana, Nebraska, Ohio, South Carolina, and West Virginia.

### **Validation: Psychometric Analyses <sup>4</sup>**

Findings of classical item statistics which indicate the overall difficulty of the ELDA items and tasks suggest that the Listening and Speaking assessments were relatively easy for students in the field test (i.e., item difficulties in the range .70 to .81) and the item difficulties for the Reading and Writing items and tasks are in more typical ranges (i.e., .54-.67). Items and tasks in the Reading, Listening, and Writing assessments are moderate to strong (item-total correlations in the range .47 to .62 range) and strong for the Speaking assessment (i.e., item-total correlations range from .81 to .87). (See Appendix B Table 1).

Rates at which examinees do not respond to test items also are relevant to the difficulty of the items and provide some indication of the level of motivation that examinees displayed on the 2004 ELDA field test. Results indicate that students omitted few items in Reading, Listening, and Writing. These rates compare favorably to non-response rates of native English speakers in academic content area assessments. The non-response rates are particularly low in the Writing assessment, which contains short and extended constructed response items, which may be omitted by as many as 5% of examinees in academic content assessments. The non-response rates in Speaking are high, particularly in the grades 6-8 cluster. These rates may suggest a range of concerns about assessing English proficiency on English language learners (e.g., reticence in assessment situations), the delivery system for the ELDA Speaking assessment (i.e., prerecorded tasks delivered via audio recording, examinees record responses for subsequent scoring), the design of the ELDA Speaking tasks (e.g., the scaffolded prompts), or the difficulty and appropriateness of the tasks themselves (e.g., the degree to

---

<sup>4</sup> For full report go to: [www.ccsso.org/projects/elda/researchh-studies](http://www.ccsso.org/projects/elda/researchh-studies).

which the tasks offer opportunity for response for the diversity of English language learners who participated in the field test). (See Appendix B Table 2).

Differential item functioning (DIF) indicates whether items function differently for examinees of equal proficiency from different subgroups. If items are unequal in difficulty for examinees of equal proficiency from different subgroups DIF, they function differently for the subgroups. This difference is considered unfair to the subgroup for which an item is relatively more difficult. DIF is relevant to the validity of inferences from test performance about examinee's English proficiency.

Results show that relatively few items were flagged for DIF in all ELDA assessments and grade clusters except in Reading and Listening grades 6-8 and Speaking grades 9-12. The LEP-SCASS reviewed all flagged items, suspended from subsequent use a small numbers of flagged items, and approved all other flagged items for subsequent use on operational test forms because they could find no content of contextual topics or features in the flagged items to explain the DIF flags and warrant discontinuing their use. (See Appendix B Table 3).

The degree to which the ELDA test forms yield scores that are free of error are indicated using an internal consistency reliability estimate, coefficient alpha. The reliability estimates for all ELDA field test forms exceed .85, except for the Writing assessments. The Writing assessments are relatively short (i.e., 19 items for 28 points in the grades 3-5 and 6-8 assessments, 20 items for 31 points in the grades 9-12 assessment) and contain a variety of items types (i.e., multiple choice and short and extended constructed responses items) that assess a range of writing skills (e.g., writing a draft,

editing). These features explain the relatively low internal consistency reliability estimates for the Writing assessments. (See Appendix B Table 4).

AIR used Masters' Partial Credit Model (1982), an extension of the one parameter Rasch model that allows for both multiple choice and constructed response items, and widely used Winsteps software to estimate item parameters for the ELDA assessments. Because each ELDA assessment (i.e., the Listening, Reading, Writing, and Speaking assessments) contains a common set of items between adjacent grade clusters, the grades 3-5, 6-8, and 9-12 forms in each ELDA area were jointly calibrated in a single Winsteps run for each subject. The joint calibration produced a common, vertically linked scale across grade clusters for each content area. For each Winsteps run, the mean of the item difficulty parameters was fixed to zero so that operational form 1 had an average difficulty (i.e., average item step value) equal to 0.0. (See Appendix B Table 5).

We examined items that Winsteps flags for misfit to the Partial Credit/one parameter model. Misfit statistics indicate items that assess language and other proficiencies that may be related but tangential to the target construct for ELDA assessment, proficiency in reading, listening, writing, or speaking. Results indicate the 1-24% of items were flagged for misfit. The LEP-SCASS reviewed all items flagged for misfit, suspended from subsequent use a small numbers of flagged items, and approved all other flagged items for subsequent use on operational test forms because they could find no features in the flagged items to explain the misfit flags and warrant discontinuing their use.

### **Validation: Validity Studies<sup>5</sup>**

CCSSO's LEP-SCASS technical advisory committee, the Center for the Study of Assessment Validity and Evaluation (C-SAVE), and AIR developed a validity research agenda. Two general types of analyses were performed by C-SAVE, item level and test level analyses, with several analyses conducted in each category to provide forms of evidence.<sup>6</sup>

The purpose of the item analyses was to facilitate CCSSO and AIR in the process of selecting items from the pool of field tested items to form valid ELDA test forms. These analyses supplemented AIR's traditional item analyses that focused on scoring keys and rubrics, item difficulty assessments, biserial and point biserial discrimination indices, and DIF.<sup>7</sup>

*Latent Class Analyses.* The main purpose of the ELDA field test was to evaluate the initial pool of test items. A different collection of items was constructed for each domain (reading, writing, speaking, and listening) for each grade cluster (3-5, 6-8, and 9-12) of students. Each collection was assigned to one of two field-test forms (A, B) so that each form reflected, as closely as possible, the final test blueprint. The field test data set included item responses for every item for each student together with collateral data (e.g., language acquisition level, primary language, type of ESOL program, standard demographics) on every student. Using the Winmira program, five-class models were fit to item response data from each field test form.

---

<sup>5</sup> This section is based on and contains excerpts (with permission) from Kopriva, R. (October, 2004). Field Test Validity Study Results: English Language Development Assessment. Final Report

<sup>6</sup> C-Save Center was formerly of the University of Maryland and now is housed at the University of Wisconsin.

<sup>7</sup> See the full validity report in: [www.ccsso.org/projects/elda/Research\\_Studies](http://www.ccsso.org/projects/elda/Research_Studies).

For each domain and grade cluster form, the proportion correct was estimated for each item within each latent class. To evaluate the validity of items for discriminating among the ordered latent classes, the differences in proportion correct between adjacent classes were calculated. (See Appendix B Table 6).

*Teacher Ratings of Student Proficiency.* The field test data collection included teacher assessment of each student's language proficiency in reading, writing, speaking, and listening. These took the form of a 5-point developmental scale in each domain. For each domain, these data were used to group students by level and calculate the proportion correct on each item in every form for each proficiency level. As one would expect, the proportion correct increases with proficiency level. In order to evaluate the validity of items for discriminating student proficiency levels as reported by the teachers, the differences in proportion correct between levels were calculated as was done for the latent class analysis results. (See Appendix B Table 7).

*Developmental Level Ratings of Items.* To analyze developmental level ratings of items experts with extensive expertise in ESOL and language testing were trained and charged with assigning to each item the performance level designation (that is, beginning, lower intermediate, and so on) that best identified the level of development at which the item was focused. Three sources of information were available to use as criteria for judging the items: the latent class gradient results, the student proficiency gradient results, and the developmental level ratings of items. In order to evaluate the items consistently over domains, grade spans, forms and item types, a flagging system was developed that would identify the strength or weakness of each item referenced to a set of criteria. The criteria were based on the degree to which the item discriminated at a single

location on the developmental scale and the consistency of evidence across the three sources. The results of the item reviews according to these criteria were used along with the traditional item analyses produced by AIR to determine whether each item should be considered for the operational forms of the ELDA test, or whether it should be revised or discarded.

*Results of Item Analyses.* The developmental ratings of the items found that, with the exception of speaking, all domains had field-test items representing each of the five developmental levels in each grade cluster, with fewer items, in general, at levels 1 and 5. Speaking had no items that were rated level 5. Reading items in grade cluster 9-12 received “strong” flags, indicating potentially weak items, more often than those in 3-5 or 6-8. The flagging pattern for listening items was more consistent across grades. A strong flag suggests that the item discriminates poorly or that all three sources conflict on where the item discriminates. Writing multiple-choice items received the highest proportion of “strong” flags overall, and only a few speaking items in the 9-12 cluster received “strong” flags.

#### Analyses of Relationships among Development Level of Items, Percent Correct and Teacher Ratings of Student

In addition to the analyses of individual item validity, the relationship of developmental ratings of items and teacher ratings of student proficiency to item difficulty was evaluated. This was done by developing cross tabulations of percentage correct by item developmental levels and student proficiency ratings, and by performing two-way mixed model ANOVAs to estimate how difficulties varied over the two factors.

The results of the ANOVAs, using percent correct as the outcome variable, were remarkably uniform over domains and grade clusters. In all cases, the main effects (development level and teacher rating) were significant. The interaction was also significant for almost all domains and grade clusters, except for reading 9-12A and B, Listening 3-5A and 6-8A, and Writing (MC) 9-12A and B. Although the findings were not significant, the results were not disordinal and continued to reflect the monotonic nature of the other analyses.

Across item developmental levels and across student proficiency levels (and for all domains and grade spans), item probabilities were ordered monotonically. That is, for items in developmental level category 1, the percentage correct uniformly increases by student proficiency level. Likewise, for student proficiency level 1, the probability correct decreases as items become more difficult in the higher developmental levels. This occurs over all developmental levels, for all domains, and in all grade spans. The results validated both teacher rating of student proficiency and the expert development level rating of items in terms of logical expectation about percents correct.

### **Field Test Score Analyses**

The quality of ELDA was reviewed in terms of how the field tests as a whole measured the targeted sets of latent traits inherent in the four domains. (Note that these analyses were not performed on final, operational test forms; although the results can be generalized somewhat to the operational forms, because test construction was conducted using the findings of the item analyses, the results do not apply directly to operational forms.) Three sets of analyses were conducted. Structural equations models were built and analyzed which investigated the relationships among ELDA, the Language

Assessment Scales (LAS), the Idea Proficiency Test (IPT), and teacher ratings with respect to how well they interpret the latent domains of reading, writing, speaking and listening. The underlying internal developmental structure of the ELDA test was also examined—specifically, the theoretical nature of the development of proficiency in one language for those whose primary language is another. Second, the latent class framework of the domain scores was evaluated both in terms of proficiency level and in terms of item group indicators that cut across performance levels. Finally, other measures of proficiency were evaluated in reference to the judgment valuations of the complexity of skills measured in the ELDA items.

### **Relationship of ELDA with Other Measures**

The relationships of ELDA scores with other measures of English language proficiency—LAS proficiency levels, IPT proficiency levels, and teacher ratings of student proficiency—were investigated for the whole group of students. Relationships were also investigated for critical subgroups identified by the SCASS: language proficiency level, including post-ESL and native English-speaking students; language/linguistic group; type of ESOL instruction; and grade level. For each grade cluster, the resulting multitrait-multimethod matrix was represented as a path model. Four latent traits were included in this model to represent the true scores on the reading, writing, listening, and speaking traits. The other latent variables (LAS, IPT, and ELDA proficiency levels and teacher ratings) were included to represent the effects of the methods.

Overall, the findings suggested convergent validity of methods across domains and some evidence of discriminant validity. The ELDA tests, LAS, IPT, and teacher

ratings were all measuring language proficiency, but, in all clusters (especially at 6-8), there appeared to be only limited ability to discriminate the domains within the measurement of language development. This remains a substantive question—how much unique variance within each domain should one expect to build into an assessment of language proficiency?

When models were fit for each group within the four subgroup categories, with very few exceptions, the ELDA tests and teacher rating score-trait correlations were higher than either LAS or IPT. For the most part, ELDA behaved very similarly to teacher ratings, while LAS and IPT loadings tended to be analogous. In general, ELDA loadings were respectable in size and stable across domains tested, suggesting stability over most groups within each of the subgroups. They also clearly differentiated ELDA findings from the other tests, which in turn tended to consistently produce substantially lower score-trait correlations.

#### Latent Class Analyses of Field Test Scores

Total test scores in each domain were evaluated via a standard latent class analysis and the mixed Rasch model. The framework for the latent class analyses is the theoretical view of English language development in which an English language learner passes through multiple stages of development, from pre-production to advanced fluency, in each of four major modes—listening, speaking, reading and writing—that are reflected in the four domains assessed by the ELDA. These separate stages are interdependent in that, e.g., listening must be at least partially developed before speaking, reading, or writing can be initiated.

The standard latent class analyses performed on field test forms were generally consistent with an ordered five-class model that captured the developmental stages of the English language development process. In contrast, the mixed-Rasch latent class method resulted in five classes for approximately 60% of the domain/grade cluster/form combinations. Lack of fit was particularly evident in the speaking domain, where no field-test form supported five classes, and in writing, where only two forms supported five classes. It is unclear whether this finding results from the developmental process being less refined for writing and speaking than for reading and listening, whether the field-test forms for writing and speaking were not sufficiently valid to allow five-stage discrimination, or whether speaking and writing are multidimensional, as measured by the field-test forms.

#### Analyses of Developmental Level Structure

The complexity of skills analyzed by the items was analyzed using a simplex structural model, based on the expert judgment valuations of the items with regard to complexity of skills. Second, the complexity of the structure of the items was evaluated relative to IPT and LAS proficiency level scores.

Using the ELDA field test item results, items were identified by developmental levels, as defined by complexity of skills required, by expert judges (see item analysis section). Because the number of items in the highest and lowest groups was too small to generate a stable score, items were grouped into three categories of development—rating levels 1 and 2 were identified as low English proficiency, level 3 formed the medium developmental category, and items in levels 4 and 5 were assumed to be those that discriminated primarily at the high level of proficiency. Mean percentage correct scores

in each of the three developmental categories were computed for each student and formed the basis of the analyses.

The data were assessed using a developmental model representing a simplex structure, via a structural equation model to fit recursive regressions for the various grade clusters and domains. The simplex structure assumption posits that skills for the most part are cumulative—more complex skills build on simpler skills for most language constructs. Overall, the results supported this hypothesis over grade clusters: Reading and listening models generally indicated a good to excellent fit, speaking and writing constructed response models suggested an adequate to good fit, and writing multiple choice findings indicated mixed but typically supportive models of fit. With some exceptions, the mean percentage correct monotonically decreased with complexity of skills being measured. Importantly, residual correlations between non-adjacent categories of complexity tended to be fairly low over grade clusters and domains, generally supporting the simplex notion.

#### Regressing Other Measures on Developmental Level Scores

One of the primary purposes of the ELDA test is to measure complex academic language proficiency skills in addition to the less complex skills addressed in more basic academic situations and in social language competency. Given the confirmation of this structure in the simplex analyses, other measures of proficiency, the LAS and IPT language proficiency levels, were regressed on the complexity of skills as defined in the expert judgment complexity valuation of ELDA scores.

The regressions of LAS and IPT on these complexity categories found that ELDA and the other dependent measures were measuring considerably different skills,

especially for writing, listening and speaking. In many cases, marginal amounts of information about skills in the LAS or IPT can be predicted from our understanding of complexity, as operationalized by ELDA. Most elusive overall is the ability to predict higher level complexity skills in the other measures. This finding is consistent with the evaluations of commercially-available tests in the literature and by the ELDA development committees, where one of the main goals of ELDA was to measure a broader range of skills, particularly higher academic proficiency skills, than the tests that were currently on the market.

### **Assembling Operational Test Forms for Administration in 2005-2007 and Beyond**

ELDA content specialists at AIR assembled draft versions of operational forms 1 (i.e., after the 2004 field test analyses) and operational forms 2 and 3 (i.e., after the 2005 field test analyses) in all content areas and grades. Each draft form underwent three levels of review by other ELDA content specialists, the ELDA test development leader, and the ELDA project director. During each phase of review, these assessment specialists worked with AIR psychometricians to ensure that each form balanced the content and statistical requirements in the ELDA specifications. The LEP-SCASS reviewed and approved operational forms 1, 2, and 3.

AIR content specialists assembled forms to meet the following specifications:

- The specified numbers of vertical linking items (i.e., common items across grade cluster assessments) in operational form 1, horizontal linking items in operational forms 2 and 3 (i.e., common items across the same grade clusters in all three forms)

- Test blueprint features, such as the number of items per standard on each test form as a whole, the balance of benchmarks and topic areas (e.g., Mathematics, Science, and Technology; School-Social Environment), and the maintenance of item type order, as indicated in the ELDA specifications
- Miscellaneous features, such as multiple choice item key counts and balance, passage topic balance, gender balance, item and classification soundness, and content overlap

Finally, items in all forms were sequenced to be consistent with the relative position of linking items in operational form 1 and the field test position of all other items being used operationally for the first time in operational forms 2 and 3. These requirements reviewed for each form and across forms in grade clusters and content areas to ensure that all assembled forms were as parallel as possible from a content and statistical perspective.

### **Standard Setting Process**

At the initial meeting of the steering committee, December 2002, members states in consultation with AIR staff, decided on five ELP performance levels. Extensive discussion took place regarding the objectives of the assessment to be developed, the breath of content coverage to be assessed, the linguistic demands of the content area under consideration, and the number of items needed in order to cover the standards and lastly the amount of time it would take to complete each of the domains tested. As the development evolved AIR and steering committee members' further refined the performance level descriptors that were instrumental in the development of the performance standards.

Performance standards for ELDA for grades 3-12 were set in August 2005. For the Reading, Writing, and Listening tests, MI used a bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001). In this procedure, standard setters evaluated specially formatted test booklets and placed bookmarks at points where the difficulty of items appeared to change in ways that differentiated between adjacent performance levels (i.e., Pre-functional, Beginning, Intermediate, Advanced, and Fully English Proficient). For the Speaking test, MI used a generalized holistic approach (Cizek & Bunch, 2007). In this procedure, standard setters evaluated live samples of student work, placing them into one of the five categories (Pre-functional to Fully English Proficient).

Standard setters worked in grade-cluster groups (3-5, 6-8, and 9-12) to set standards for all tests in Reading, Writing, and Listening. A single group set standards for all Speaking tests. At the close of a three-day standard-setting meeting, the individual groups turned their recommendations over to an Articulation Committee composed of representatives of each of the four initial groups. The function of the Articulation Committee was to merge the individual grade-cluster performance standards into a set of standards that would span the grades, eliminating or smoothing any cluster-to-cluster disparities or discontinuities they might find. The Articulation Committee also recommended procedures for combining scores to produce Comprehension and Composite scores. All cut scores were subject to final review and approval by CCSSO.

## **ELDA K-2 Assessment**

### **Theoretical Basis**

The design of the K-2 ELDA is informed by current view of early childhood development and implications for assessment. Specifically, that at this age children grow

and change rapidly in terms of their motor, language, cognitive and social-emotional development. Consequently, in the development of the assessment special attention was devoted to the overall time of the student observation, the format of the assessment, the interaction between teacher and student, the supports available to teachers and students [pictures, manipulative], and the complexity of the language of the prompt. Given the dearth of empirical data regarding best practices of assessing young ELL learners, the member states and test developer relied on the work focusing on second language learning at this age and instructional practices appropriate for young ELL learners. This effort was coupled with on-going input from expert consultants and EC teachers from member states.

### **Defining/Using State Content Standards**

The process for incorporating state content standards into test specifications for K-2 was identical to that used for 3-12 (cf. American Institutes for Research, 2003). In November 2003 members of the K-2 advisory sub-committee of met with staff from AIR to review state ELP content standards and select those appropriate to K-2. Members and CCSSO staff combined identical or similar content standards from member states eliminated those whose evaluation would be beyond the scope of the proposed methodology of the assessment, and prepared a final, consolidated set of content standards.

The condensed standards and K-2 assessment framework were generated by AIR with support from early childhood education consultants and K-2 subcommittee members of the LEP SCASS states. The standards accepted by the membership were subsequently reviewed by MI development staff in preparation for item development.

## **Test Blueprint and Item Development**

Early in the development process state members opted for an inventory approach over the traditional multiple-choice (MC) and constructed-response (CR) item approach because of the age and developmental stage of the student population (kindergarten through second grade). Each "item" in the initial inventories was a statement regarding an observable student behavior such as the following:

- Follows a two-step instruction in a non-academic setting (e.g., going to the lunchroom)
- Identifies a picture of an object with the same ending sound as 'cat'
- Uses correct English words for manipulatives content, age, and grade appropriate

MI invited nine classroom teachers to participate in an item development session in Durham, NC in February, 2005. The teachers, who were drawn from member states, worked with MI staff and chair of the LEP SCASS K-2 subcommittee to create inventory entries for Listening, Reading, Speaking, and Writing. The work was collaborative and had as its goal the generation of enough teacher observations/ items to construct three field test inventories.

The item writers used the lists of standards and benchmarks collected from the LEP SCASS member states in the consortium as their guides, along with other materials they brought to the session and those supplied by MI.

The final step in item development was the selection of anchor items from the current tests for grades 3-5. These items were selected in order to link scores of k-2 assessments to those of the assessments for grades 3-12. These items were selected on the basis of their relevance to the K-2 content objectives and the fact that they were

among the easiest of the ELDA 3-5 items, suggesting that they would not be too difficult for those students in grades 1-2.

### **Field Testing**

Six states (Indiana, Kentucky, Nebraska, New Jersey, Oklahoma, and West Virginia) participated in the field test, with a total of 2,431 students (745 K, 831 grade 1, and 798 grade 2). MI scoring leaders conducted training for scorers as they did in 2004, and those scorers evaluated student responses to the writing prompts as well as responses to the speaking prompts.

### **Reliability and Validity**

The K-2 inventories were administered in the spring of 2005 in their preliminary (long) version. Results are documented in CCSSO (2006) and summarized here.

*Item face validity.* MI staff, CCSSO staff, and two nationally recognized content experts met for a face-to-face review session in CCSSO's offices in Washington, DC on March 11, 2005. This session was similar to those conducted in 2004 for items developed for grades 3-12. At the end of the review session, MI staff documented all recommendations, made the necessary modifications, and submitted all items to CCSSO for final approval. The basic structure of the inventories was validated by the two content experts, who provided suggestions for refocusing specific inventory entries (items) and approved the instruments for field testing.

*Item reliability.* Corrected item/total correlations for all inventory rows (items) ranged from .48 to .87, indicating an extremely high internal consistency as measured at the item level.

*Item response vs. teacher rating.* The same analyses revealed correlations between item scores and teacher ratings ranging from .24 to .65, with most (60 out of 63) above .3, and 40 out of 63 had correlations above .5.

*Item response by grade.* The technical report includes analyses of item response by grade. With one exception, all grade-to-grade differences in item score were positive (Speaking item 12 had a difference of 0 from K to grade 1). In general, differences in Reading and Writing were much higher (a total of a full point, on average from K to grade 2 in Reading and just under a full point from K to grade 2 in Writing) than in Listening and Speaking (about a quarter of a point from K to grade 1 for both and just under half a point from K to grade 2 in both). Overall, however, the indication was that all but one item showed gains from grade to grade.

*Test reliability.* Generalizability analyses showed the inventories to have reliability coefficients ranging from .92 (Listening, 7 rows) to .97 (Reading, 29 rows). A reliability coefficient of .90 is considered to be excellent for individual decisions about students.

*Test score vs. teacher rating.* Correlations between inventory scores and teacher ratings of student proficiency ranged from .57 (Listening) to .68 (Reading and Speaking). The correlation for Writing was .58. All of these correlations reveal a strong relationship between scores on the inventories and classroom teacher judgments about students' levels of proficiency.

### **Operational Form Results**

In the spring of 2006, ELDA K-2 (shortened version) was administered to 21,228 students in four states. Analyses were similar to those performed in 2005.

*Item reliability.* Corrected item/total correlations for all inventory rows (items) ranged from .58 to .86, slightly higher than in the field test, again indicating an extremely high internal consistency as measured at the item level.

*Item response vs. teacher rating.* The same analyses revealed correlations between item scores and teacher ratings ranging from .45 to .77.

*Item response by grade.* For the operational forms, the entries for the kindergarten level are different than those for grades 1-2; therefore direct comparisons are available only for grades 1-2. All differences were positive, ranging from .10 (Reading, item 1) to .47 (Reading item 4), with a mean of about one-fifth of a point from grade 1 to grade 2 for a given item.

*Test reliability.* Even though all inventories except Listening were considerably shortened, (Reading, for example, from 29 rows to 14), all reliability coefficients were above .90. Test reliability ranged from .94 (Listening, all grades) to .97 (Reading, grade 1). Given this range, it is safe to conclude that there was little variability at all in total test reliability, and that the predictions based on the field test results were quite accurate.

*Test score vs. teacher rating.* Correlations between inventory scores and teacher ratings of student proficiency ranged from .65 (Writing, kindergarten) to .77 (Reading grades 2 and 3). All correlations reveal a strong relationship between scores on the inventories and classroom teacher judgments about students' levels of proficiency.

### **Standard Setting Process**

Performance standards for ELDA K-2 were based on performance level descriptors developed specifically for these assessments by Malagon, Rosenberg, & Winter (2005). For K-2, the holistic approach was used for all inventories. Performance

standards were set in a web conference in January 2006 and confirmed at a second web conference in July 2006. Details of conferences, procedures, and outcomes, are described in Bunch & Joldersma (2006).

### **Creating Operational Forms**

Subsequent to the 2005 field test, there were two key meetings concerning ELDA K-2. The first was in Savannah, Georgia, on July 6-7, 2005. At this meeting, state representatives presented many of the concerns voiced by K-2 teachers regarding the length and difficulty of administering the inventories. In August, Dr. Dina Castro joined the MI team of developers to begin revising the inventories with two key goals: shorten the inventories, and provide more support for teachers who administer them. On December 7-9, 2005, member state representatives met again in Washington, DC, to review revised materials. These materials were ultimately approved with modifications during December 2005 and January 2006. The final materials were submitted and approved on January 31, 2006.

### **Accommodations and Validity (K-12)**

The administration manual of the ELDA assessment sets forth guidelines for offering and using accommodation for ELL students with disabilities. The recommendations were informed by members understanding of special education requirements and extensive consultations with knowledgeable experts. Generally, the guidelines recommends that accommodations should always be related to the student's specific disability, and that they be consistent with those allowed in the students IEP or 504 plan and with practices routinely used in the student's instruction and assessment. Since ELDA is a language assessment, and most accommodations offered to ELL

students are language related, only certain types of assessments are recommended with the ELDA: computerized Assessment; dictation of Responses; extended/adjusted time; and individual/small group. There was recognition among members states that the research evidence for the use of these accommodations is limited and that more research is needed on the validity of these accommodations. Finally, in addition to those listed above, Braille and large print versions of the Reading and Writing ELDA are permitted.

For the grades 3-12 ELDA in 2005 and 2006, large print and Braille versions of the Reading and Writing tests were produced and shipped as requested by schools. The Listening and Speaking tests were not produced in large print and Braille formats because the students respond to audible stimuli for these tests. In addition to modified test booklets, other permissible accommodations for ELDA testing included computerized assessment (typing of open-ended writing responses), extended/adjusted time for completion of the assessment, individual/small group administration, dictation of responses for all parts of the assessment with the exception of the constructed response Writing items, and any accommodations provided for under an individual student's documented IEP or 504 plan.

### **Test Administration and Technical Manual (K-12)**

The grades 3-12 ELDA was designed to be administered to class-sized groups of students simultaneously, with the exception of the Speaking assessment. This assessment is scored live on site by teachers. In 2005 and 2006, separate Speaking Scoring Guides were developed containing text of the audible questions that students respond to orally, and sample answers to each question representing each score point. The teachers then

bubbled in the student's score to each question on his or her individual answer document, and the scores were captured.

The grades K-2 assessment consisted entirely of inventory items completed by teachers. Teachers recorded the scores for each item in each student's test booklet, guided by information in the 2005 K-2 Test Administration Manual, and in 2006, by the Test Administration Manual and Teacher Support Materials. Because the inventories differ between grade K and grades 1-2, two versions of the Teacher Support Materials were developed.

At the conclusion of testing in 2005 and in 2006, AIR and MI collaborated to produce Technical Manuals. These manuals (AIR, 2005; CCSSO, 2006) describe test development, administration, scoring, analyses, and outcomes in some detail.

### **Scoring and Reporting (K-12)**

Because K-2 inventory scores are recorded directly into student test booklets and not onto scannable forms, trained operators record the inventory scores in a tested Data Entry application. The ESBQ for each student is then scanned, and the demographic data is captured and stored in databases divided by state. The scores are then merged with the demographic data via a unique matching bar code on each student's test booklet and ESBQ.

For ELDA K-2, MI staff key entered identification information and scores from the inventories to a data file using a double-entry procedure. Entries are post-edited for out-of-range entries and other anomalies prior to uploading to score reporting programs. Score reporting is the same as for ELDA 3-12 versions.

For grades 3-12 trained readers scored the constructed response Writing items according to initial rangefinding results. CCSSO conducted a rangefinding with participants from ELDA consortium states in 2004. Supervised readers assigned scores based on the scoring protocols determined at rangefinding, and then bubbled in their scores on scannable scoring monitors. Ten percent of all responses received a second reading to verify reliability of readers' scores. The monitors were then scanned, and the results were merged with the data derived from the students' multiple choice answer selections.

After all scores were processed, MI created the following PDF files for each district:

- Demographic Report
- District Summary Report
- Student Roster
- Individual Student Report
- Student level data file

The PDF files for each district are then transferred to a CD and sent to each district. Each state also received a CD with their corresponding files. The data sets included not only raw scores and levels but scale scores, school and district information, student demographic data, and program information. Score reports included an interpretive section which explained the five performance levels, provided scale score ranges for all performance levels for all tests, and described how the Comprehension and Composite scores were derived.

## References

- American Institutes for Research (2005a). English Language Development Assessment Test Specifications and Standards Document. Washington, DC: Author.
- American Institutes for Research (2005b). English Language Development Assessment (ELDA) Technical Report: 2005 Operational and Field Test Administration. Washington, DC: Author.
- AIR/CCSSO/LEP-SCASS (undated). English Language Development Assessment K – 2 Standards and Benchmarks. Washington, DC: AIR.
- Brennan, R. L. (1983). Elements of Generalizability. Iowa City, IA: American College Testing Program.
- Bunch, M. B. (2006). ELDA Standard Setting Final Report. Durham, NC: Measurement Incorporated.
- Bunch, M. B. & Joldersma, K. (2006). Setting Standards for ELDA K-2. Durham, NC: Measurement Incorporated.
- Butler, F. A., Lord, C., Steven, R., Borrego, M., & Bailey, A. L. (2004 April). *An Approach to Operationalizing Language for Language Test Development Purposes: Evidence from Fifth-Grade Science and Math*. (CSE Report 626.) Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

In Abedi, J. (2007). English Language Proficiency Assessment in the Nation: Current Status and Future Practice (Ed.): Davis" University of California.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

Council of Chief State School Officers (2006). English Language Development Assessment K-2 Technical Manual: Spring 2006. Washington, DC: Author.

Cummins, J. (1979) *Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters*. Working Papers on Bilingualism, No. 19, 121-129.

Malagón, M. H., Rosenberg, M. B., & Winter, P. (2005). Developing Aligned Performance Level Descriptors for the English Language Development Assessment K-2 Inventories. Washington, D.C.: CCSSO.

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

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Editor), *Setting Performance: Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum.

**Appendix A: Members of the ELDA Initial Steering Committee [December, 2002]**

<b>Jamal Abedi</b>	CRESST/UCLA
<b>Cori Alston</b>	South Carolina Department of Education
<b>Rebecca Blum Martinez</b>	University of New Mexico
<b>Barbara Carolino</b>	Council of Chief State School Officers
<b>Eduardo Cascallar</b>	American Institutes for Research
<b>Michael Fast</b>	American Institutes for Research
<b>Steve Ferrara</b>	American Institutes for Research
<b>Guillermo Solano Flores</b>	WestEd
<b>Rebecca Kopriva</b>	University of Maryland
<b>Julia Lara</b>	Council of Chief State School Officers
<b>Leslie Lightbourne</b>	Louisiana Department of Education
<b>Carlos Martinez</b>	U.S. Department of Education
<b>Mary Sue Morin</b>	Nevada Department of Education
<b>John Olson</b>	Council of Chief State School Officers
<b>Robin Scarcella</b>	University of California, Irvine
<b>Roberta Schlicher</b>	Virginia Department of Education
<b>Maria Seidner</b>	Texas Education Agency
<b>Carmen Sosa</b>	Iowa Department of Education
<b>Jeanette Spencer</b>	California Department of Education
<b>Kelly Westphalen</b>	American Institutes for Research
<b>Lily Wong-Fillmore</b>	University of California, Berkeley

### Appendix B: Tables

**Table 1. Mean Item Difficulty and Discrimination Statistics**

	<b>Item Difficulty</b>	<b>Item Discrimination</b>
Reading	.61-.67	.56-.60
Listening	.70-.72	.60-.62
Speaking	.77-.81	.81-.87
Writing	.54-.59	.47-.53
<i>Note.</i> Ranges of means across forms and grade clusters in the 2005 field test forms.		

**Table 2. Non-Response Rates (in Percentages)**

	<b>Grade Cluster</b>		
	<b>3-5</b>	<b>6-8</b>	<b>9-12</b>
Reading	1.8	1.3	2.3
Listening	0.3	0.5	4.3
Speaking	4.7	12.0	7.1
Writing	0.6	1.2	1.2
<i>Note.</i> Across 2005 field test forms and grade clusters. Combination of items skipped and items not reached.			

**Table 3. Items Flagged for Differential Item Functioning (DIF)**

	<b>Grade Cluster</b>		
	<b>3-5</b>	<b>6-8</b>	<b>9-12</b>
Reading	9/162 (6)	25/168 (15)	18/192 (9)
Listening	7/150 (5)	15/150 (10)	12/180 (7)
Speaking	3/60 (5)	0/60 (0)	18/60 (30)
Writing	4/76 (5)	3/76 (4)	1/80 (1)
<i>Note.</i> Across 2005 field test forms within grade cluster. Comparisons are made for males vs. females, speakers of Spanish vs. other foreign languages, and students currently in limited English proficiency programs vs. students exited from such programs. Numbers of items flagged/total number of items; percentages in parentheses.			

**Table 4. ELDA Score Reliabilities: Coefficient Alpha**

	Grade Cluster		
	3-5	6-8	9-12
Reading	.93	.93-.94	.94-.95
Listening	.91-.92	.92-.93	.94-.95
Speaking	.88-.90	.93-.94	.88-.92
Writing	.76-.82	.84-.85	.84-.87

*Note.* Across 2005 field test forms within grade cluster.

**Table 5. Items Flagged for Misfit in IRT Calibrations**

	Grade Cluster		
	3-5	6-8	9-12
Reading	33/162 (20)	27/168 (16)	32/192 (17)
Listening	36/150 (24)	34/150 (23)	35/180 (19)
Speaking	8/60 (13)	12/60 (20)	11/60 (18)
Writing	1/76 (1)	13/76 (17)	3/80 (4)

*Note.* Across 2005 field test forms within grade cluster. Numbers of items flagged/total number of items; percentages in parentheses.

**Table 6. Latent Class Analysis**

Item Order	Class A	Class B	Class C	Class D	Class E
	1	0.51	0.87	0.96	0.98
2	0.62	0.98	1.00	0.98	1.00
3	0.63	0.98	1.00	0.99	1.00
4	0.45	0.78	0.88	0.93	0.97
5	0.35	0.61	0.85	0.95	0.97
6	0.22	0.48	0.76	0.83	0.93
7	0.39	0.86	0.97	0.98	0.99
8	0.25	0.73	0.94	0.97	0.99
9	0.41	0.62	0.87	0.94	0.98
10	0.51	0.84	0.98	0.99	0.98
11	0.28	0.65	0.91	0.98	0.99
12	0.26	0.64	0.93	0.96	0.98
13	0.16	0.20	0.33	0.50	0.61

**Table 7. Proportion Correct by Student Proficiency Level**

Item Order	<u>Student Proficiency Ratings (PR)</u>				
	1	2	3	4	5
1	0.55	0.87	0.95	0.98	1.00
2	0.68	0.96	0.98	0.99	1.00
3	0.68	0.97	0.98	1.00	0.99
4	0.48	0.82	0.89	0.91	0.93
5	0.41	0.67	0.86	0.89	0.97
6	0.25	0.59	0.74	0.78	0.90
7	0.48	0.87	0.94	0.97	1.00
8	0.42	0.79	0.90	0.92	0.97
9	0.47	0.68	0.88	0.92	0.94
10	0.64	0.86	0.95	0.96	0.99
11	0.36	0.77	0.87	0.93	0.97
12	0.38	0.72	0.88	0.93	0.96
13	0.18	0.25	0.36	0.43	0.58