



*Louisiana Believes.*

## **English Language Development Assessment**

### **2013 Technical Summary**

The tests used in Louisiana are constructed to fairly assess the progress of Louisiana students. As such, the development process and statistical or psychometric work are carried out with great care. This document provides an overview of the process and summarizes some of the key psychometric information.

#### ***Introduction***

The *No Child Left Behind Act* (NCLB) of 2001 requires states to annually assess English proficiency in listening, speaking, reading, writing, and comprehension and to report annual progress or attainment of English proficiency for all students identified as limited English proficient (LEP) in kindergarten through grade 12. In December 2002, sixteen states formed a consortium under the coordination of the Council of Chief State School Officers (CCSSO) to begin the development of a criterion-referenced English language proficiency assessment that would meet the NCLB requirements. The English Language Development Assessment (ELDA) is aligned to Louisiana’s English language development standards and is composed of tests for four grade clusters (K–2, 3–5, 6–8, and 9–12) in four language domains (Listening, Speaking, Reading, and Writing). It assesses both the academic and school/social environment language of students. ELDA is vertically linked across grade clusters and has five levels of performance descriptors, ranging from level 1, which has a realistic definition of English proficiency for beginners, to level 5, which has a rigorous definition of full English proficiency.

Beginning in the 2012–2013 academic year, Louisiana will administer the shortened version of ELDA. In 2009, the states administering ELDA began researching the possibility of shortening ELDA while still maintaining high reliability, validity, and content coverage. Based on the research, field testing, and teacher feedback, Reading, Listening, and Speaking tests for grades 3 through 12 were shortened. In addition, Composite scores were calculated for grades 3–12 and are included on Student Reports and Student Labels. The kindergarten inventory, grades 1 and 2 inventory, and the Writing assessment for students in grades 3 through 12 were not shortened.

The CCSSO provided the leadership for ELDA to meet the requirements of NCLB. Test development work for grades 3 through 12 was done under the auspices of the American Institutes for Research (AIR). The Center for Study of Assessment Validity and Evaluation (C-SAVE) at the University of Maryland conducted the test validity research. Measurement Incorporated (MI) administered the field test for grades 3 through 12 in 2004, conducted

standard setting in summer 2004 and 2005, and provided project management for testing in spring 2005. MI also developed ELDA for grades K through 2 and field tested the K–2 assessment in spring 2005. Beginning with the spring 2006 administration, Data Recognition Corporation provides project management and all scoring and reporting for Louisiana.

Beginning in spring 2006, Louisiana administered ELDA in all grade clusters, K–2, 3–5, 6–8, and 9–12. This technical report provides item- and form-level results from the 2013 administration.

### *Development Process*

ELDA was aligned to state English language proficiency (ELP) standards through an analysis of the ELP standards of consortium states available to the project at the outset. From an analysis of state ELP standards for each of the four skills domains, AIR constructed and the LEP-SCASS approved a set of core ELP standards, which formed the basis for item design.

To develop items that measure these language proficiency standards as specified by the content specifications, AIR brought together a highly competent pool of item writers, using a mix of external item writers, NAEP foreign language item writers, and other internal content experts. AIR staff, and assessment development consultants, working in groups by domain and grade level, trained the item writers by explaining general item writing principles, and helped the participants develop items. The items and prompts were written in the language of the classroom and of the academic subjects. After items were drafted and reviewed by the writers, the following review levels were conducted:

- Preliminary: Items reviewed by junior staff for formatting and basic item construction principles
- LABS: Items reviewed by a trained and certified LABS (language accessibility, bias, and sensitivity) reviewer
- Editor: Items reviewed for grammar, writing conventions and clarity
- Senior: Items reviewed by a senior content expert in ESL or English language arts, evaluating the items for their match to the standards and for their measurement integrity

Items that passed all these reviews were brought to LEP-SCASS meetings for review, comments, revision, and approval. At SCASS content review meetings, members split into grade-cluster groups and were instructed on the specifications for the items, the standards and the benchmarks, and individually reviewed the items before meeting as a group to accept, revise, reject or recommend revisions for resubmission of all the items. In addition, the items, prompts, and data from field testing were reviewed for possible cultural bias. Those items that survived the final review entered the field-test item pool. More detailed information can be found in *ELDA Technical Report, 2004 Field Test Administration* (2005).

As test specifications require, ELDA incorporates multiple-choice and constructed-response items as well as graphic prompts and teacher-scored rubrics. Constructed-response items include short constructed-response (SCR), extended constructed-response (ECR) items, and spoken-response (SRI) items. During field testing, student oral responses were taped and scored at AIR.

For operational testing, student oral responses to the recorded speaking prompts are scored by the test administrator according to the Speaking scoring rubrics provided in a *Speaking Scoring Guide*. Recorded dialogues and presentations also are the basis for the Listening assessment. Supporting graphic prompts help clarify what is required of the student or motivate the student to give an appropriate response.

The ranges of possible points students can earn for the different types of items that make up ELDA are:

- multiple-choice items (MC) 0–1 points
- short constructed-response items (SCR) 0–2 points
- extended constructed-response items (ECR) 0–4 points
- spoken-response items (SRI) 0–3 points for grades K–2 and 0–2 points for grades 3–12

ELDA endeavors to measure English language skills independently of a student’s prior knowledge of particular content areas while using contextual school language. To best demonstrate student mastery of content-embedded language, each score domain incorporates approximately equal proportions of material from the following areas:

- English language arts
- mathematics, science, and technology
- social studies
- school-environment

In addition to the academic areas, the school-environment items emulate the language demands of the classroom and school environments.

### ***Equating of Test Forms***

The primary purpose of form equating is to establish score equivalency between two (or more) forms. Equivalency is established by first building the forms to be equated to tight content specifications, then placing the form scores on the same scale, such that students performing on an assessment at the same level of (underlying) achievement should receive the same scaled score, although they may not receive the same number-correct score. The raw score to scaled score relationship performs this leveling function, based on form equating studies. Differences in the raw score to scaled score relationship between the two forms can be due to differences in item difficulty and/or differences in the samples utilized for calibration.

The forms administered in grade clusters in 2013 were previously administered by the CCSSO consortium. This allowed for the previously-established raw-score-to-scaled-score tables to be applied. Please refer to *The ELDA Technical Report: 2005 Operational and Field Test Administrations* written by American Institutes for Research and Measurement Incorporated for the calibration and linking procedures.

### ***Validity***

A test is considered valid for a given purpose when the test measures what it is intended to measure (e.g., grade cluster 3–5 Listening) and the resulting scores are used to make inferences in the test’s intended measurement domain (e.g., whether a student in grade cluster 3–5 is performing at the grade cluster level in listening). Validity is not a property of a test but a function of the appropriate use of test scores and inferences made from test scores. For information regarding validity of the forms used in grade clusters 3-5, 6-8, and 9-12 see *The ELDA Technical Report: 2005 Operational and Field Test Administrations*. Validity information regarding the forms used in kindergarten and grades 1 and 2 can be found in the *ELDA K-2 Technical Manual Spring 2006*.

### ***Reliability***

Reliability describes the consistency and accuracy of the test scores. The more reliable a test, the less measurement error is associated with the test scores. Table 1 provides the number correct score statistics from the population data for the spring 2011 test administration. The test means and standard deviations are based on number-correct (NC) data. NC refers to the raw total score obtained by each student, and is used in the calculation of classical test statistics. Reliability is reported in the last two columns of the table. The traditional method, Cronbach’s alpha, is reported in the last column. Given the assumptions of this method and the characteristics of the tests, however, this method typically underestimates the reliability of the test. Because of this underestimation, a second form of reliability, the stratified alpha (Qualls, 1995), is computed. The second method considers the characteristics of the test design, namely the inclusion of constructed-response items. These items are typically scored in a graded fashion across a range of possible points.

Another important statistic that is reported in the table below is the standard error of measurement, which can be found in the column labeled SEM. The SEM is reported in number correct raw score units. It is expected that 68% of the time a student’s true score would fall within one SEM around that student’s observed score.

**Table 1 Number Correct Test-Level Summary Statistics**

Grade Cluster	Domain	Form	Number of Items	Total Score Points	Mean P-Val	NC Mean	NC Standard Deviation	NC SEM	Reliability	
									Stratified	Cronbach
K	Listening	1	7	21	0.62	12.99	5.02	1.36	NA	0.93
	Speaking	1	8	24	0.65	15.61	5.92	1.26	NA	0.95
	Reading	1	14	42	0.57	23.84	10.39	2.40	NA	0.95
	Writing	1	9	27	0.51	13.72	6.66	1.69	NA	0.94
1-2	Listening	1	7	21	0.74	15.47	4.86	1.19	NA	0.94
	Speaking	1	8	24	0.75	17.89	5.61	1.17	NA	0.96
	Reading	1	14	42	0.69	28.88	10.71	1.97	NA	0.97
	Writing	1	9	27	0.70	18.90	6.56	1.54	NA	0.95
3-5	Listening	SF2	35	35	0.75	26.30	6.36	2.16	NA	0.88
	Speaking	SF2	12	24	0.83	20.04	5.42	1.40	NA	0.93
	Reading	SF2	35	35	0.72	25.35	7.45	2.23	NA	0.91
	Writing	2	19	28	0.60	17.06	5.46	2.29	0.85	0.82
6-8	Listening	SF2	35	35	0.79	27.67	7.25	2.03	NA	0.92
	Speaking	SF2	12	24	0.81	19.46	6.66	1.30	NA	0.96
	Reading	SF2	35	35	0.71	24.69	7.99	2.25	NA	0.92
	Writing	2	19	28	0.68	18.39	6.08	2.21	0.90	0.87
9-12	Listening	SF2	35	35	0.78	27.26	7.42	2.06	NA	0.92
	Speaking	SF2	12	24	0.82	19.61	6.71	1.33	NA	0.96
	Reading	SF2	35	35	0.65	22.87	8.60	2.39	NA	0.92
	Writing	2	20	31	0.65	20.65	6.44	2.42	0.89	0.86

## ***References***

AIR and Measurement Incorporated (2005), *ELDA Technical Report: 2005 Operational and Field Test Administrations*. American Institutes for Research, Washington, D.C.

AIR and Measurement Incorporated (2005), *ELDA Technical Report: 2004 Field Test Administrations*. American Institutes for Research, Washington, D.C.

CCSSO and Measurement Incorporated (2006), *ELDA K-2 Technical Manual*. CCSSO, Washington, D.C.

Qualls, A. L. (1995). Estimating the Reliability of a Test Containing Multiple Item Formats. *Applied Measurement in Education*, 8, 111-120.