



## **2006 LAA 2 Technical Summary**

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

### ***Introduction***

LEAP Alternate Assessment, Level 2 (LAA 2), is Louisiana's criterion-referenced testing program for students with persistent academic disabilities. These tests measure how well a student has mastered the state content standards.

LAA 2 is designed for students whose Individualized Education Program (IEP) reflects a functioning grade level in English language arts (including reading) and/or mathematics at least three grade levels below the grade in which they are enrolled. The student's instructional program must be predominantly academic in nature. The LAA 2 policy, based on modified academic achievement standards, allows students with persistent academic disabilities who are served under the Individuals with Disabilities Education Improvement Act (IDEA) to participate in academic assessments that are designed to measure student learning. However, the decision to test the student in LAA 2 is not based on a disability category or placement setting and is not determined administratively. (The complete list of criteria can be found on the LEAP Alternate Assessment, Level 2 Participation Criteria form. The form is available on the LDE Web site, [www.louisianaschools.net](http://www.louisianaschools.net).) The LEAP Alternate Assessment, Level 2 Participation Criteria form must be completed annually by the IEP team for each student under consideration for LAA 2.

For spring 2006, LAA 2 was available only for eligible students in grades 4, 8, 10, and 11. The LAA 2 English language arts and mathematics tests were administered to students in grades 4, 8, and 10; science and social studies tests were administered to students in grade 11.

## *Development Process*

The LAA 2 test forms are comprised of items selected from LEAP and GEE item banks. All items were field tested and many appeared on spring administrations of those assessments. All items chosen for the LAA 2 assessment are aligned with Louisiana content standards and benchmarks and were reviewed for potential bias and sensitive material. Items were selected according to guidelines established by Louisiana educators representing special education and general education content specialists. State content committees and specialists from Louisiana's department of education collaborated to create the LAA 2 test forms.

The initial step in the development process was to define the population and content. The state content committees and specialists from Louisiana's department of education met for an information gathering session to

- identify characteristics of the range of students who would be eligible for LAA 2.
- articulate, standard-by-standard, which tasks would best differentiate students in this population who were mastering essential knowledge and skills from those who were not.
- identify on which cognitive tasks students in this group were most likely to be successful and on which tasks they would be most challenged.
- consider the various test designs, or formats, that the LAA 2 form could take.

Next, Louisiana's current pool of operational items (LEAP, GEE) in ELA and mathematics for grades 4, 8, and 10 and in science and social studies for grade 11 were reviewed. Using items, passages, and data (e.g., p-values for general education students and for special education students, when available), sample items (and, for ELA, passages), for each grade and subject area were selected. These items, both multiple-choice and constructed-response, represented a range of content standards and assessed a wide variety of skills in different ways. It was intended that the sample pools of items be useful in continuing to define, more specifically, which tasks were appropriate challenges for the ability range of LAA 2 students and which tasks were inappropriate.

Louisiana teacher groups in grade- and subject-specific groups to review items in the sample sets and draw conclusions about the appropriateness of each *type* of item, given the criteria defined during Step 1. These conclusions would help inform decision-making about test content and specifications.

Teachers were requested not to evaluate the quality of the particular item, but to think only in terms of the *item type* that it represented. LDE also informed the group that the goal was to use primarily unedited multiple-choice items from the existing pool; only constructed response items in all subjects and proofreading items in ELA were likely to be approved for revision. In completing this task, teachers were asked to consider the following questions:

- How appropriate is an item like this for the LAA 2 student?
- How likely would it be for the range of LAA 2 students to answer this type of item correctly?

- For which subgroup(s) of LAA 2 students might the challenges associated with this type of item be appropriate?
- For which subgroup(s) of LAA 2 students might the challenges be inappropriate?

For each question, teachers were asked to consider the appropriateness of the:

- content assessed (relevance, appropriateness, level of importance).
- complexity of the cognitive task required to answer the item correctly (including vocabulary, reading level, processing level, amount of information needed for processing, level of abstraction, level of inference).
- item format (including font size of text and labels, graphics dimensions and level of complexity).

Following the teacher-committee sessions, analysts began reviewing pools of operational items. Using the comments and recommendations from the committees, items were identified that appeared to meet the following criteria:

- appropriate content for the range of LAA 2 students or for a particular subgroup of LAA 2 students
- appropriate cognitive task for the range of LAA 2 students or for a particular subgroup of LAA 2 students
- appropriate item format for the range of LAA 2 students or for a particular subgroup of LAA 2 students

Attention also was given to ensure appropriate representation of the depth and breadth of GLEs across grade levels.

Modified test designs for each subject area that provided more specific guidelines for test length, MC/CR proportions, and acceptable item modifications (e.g., to items in the ELA Using Information Resources and proofreading sections). It was also decided to score the writing prompt using two of the six dimensions used in LEAP and GEE, the composing dimension and the audience awareness dimension.

### ***Measurement Model***

Given the small sample size of the LAA 2 assessments, maintaining a correspondence with the same type of IRT models used with the LEAP/GEE or iLEAP assessments was not advisable since item parameter estimates would not be estimated accurately. Instead, a Rasch partial credit measurement model (RPCM), which works better with smaller sample sizes, was used with the LAA 2 assessment. In the Rasch model only the item and category step location (difficulty) parameters are estimated during the calibration process.

The RPCM is an extension of the Rasch (one-parameter Item-Response Theory) model attributed to Georg Rasch (1980), as extended by Wright and Stone (1979), Masters (1982), Wright and Masters (1982) and Lineacre and Wright (2001).

The RPCM was selected because of its flexibility in accommodating both multiple-choice data as well as multiple-response category data and for its ability to maintain a one-to-one relationship between derived scores (i.e., scaled scores) and the underlying raw score scale. It is the underlying Rasch scale that facilitates equating of multiple test forms and allows for comparisons of student performance across years. Additionally, the underlying Rasch scale facilitates the critical maintenance of equivalent performance standards across the years. The RPCM is defined via the following mathematical measurement model where, for a given item involving  $m$  score categories, the probability of person  $n$  scoring  $x$  on prompt  $i$  is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^x (B_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (B_n - D_{ij})},$$

where,  $x = 0, 1, 2, \dots, m$ , and,

$$\sum_{j=0}^0 (B_n - D_{ij}) = 0$$

The RPCM provides the probability of a student scoring  $x$  on the  $m_i$  step of question/prompt  $i$  as a function of the student's proficiency level  $B_n$  (i.e., sometimes referred to as 'ability') and the step difficulties ( $D_{ij}$ ) of the  $m$  steps in prompt  $i$  (See Masters, 1982, for an example). Note that for multiple-choice questions there are only two score categories: (a) 0 for incorrect response, and (b) 1 for a correct response, in which case the RPCM reduces to the standard Rasch one-parameter IRT model and the resulting single step difficulty is more properly referred to as an item difficulty.

Under the RPCM scaling, all items, independent of type (multiple-choice or constructed-response) are placed onto a same common score scale. At the conclusion of the item parameter estimation (calibrations), all item and task difficulty estimates as well as all student proficiency level estimates are directly comparable on the common underlying logistic scale.

To estimate the parameters of the RPCM model described above WINSTEPS (version 3.54) parameter estimation software was used. WINSTEPS employs a joint maximum likelihood approach to estimation (JMLE), which jointly estimates both the person and item parameters. Parameters are jointly estimated for both the multiple-choice and constructed-response items simultaneously, placing the results onto a common RPCM scale. WINSTEPS is the most widely used parameter estimation software for use with a variety of Rasch models.

### ***Establishing Achievement Levels***

LEAP and GEE item parameters for LAA 2 items were used to project how the LAA 2 tests might align with the LEAP/GEE score-scale and/or performance standards. The details of the methodology of projecting the LEAP/GEE achievement level cuts onto the LAA 2 test are provided in a separate document *Estimating the Correspondence of the LEAP Alternate Assessment, Level 2 (LAA 2) tests with the LEAP/GEE Performance Standards*.

Based on the results of the aforementioned study, a reasonable correspondence was established between the new LAA 2 tests and two key achievement level cuts on the LEAP/GEE tests (*Approaching Basic* and *Basic*).

LDE and its technical advisors determined that four achievement level categories were most appropriate for the LAA 2 assessment. The four LAA 2 achievement levels would be organized as follows:

<b>LAA 2 Achievement Level</b>	<b>Corresponding LEAP/GEE Level</b>
<i>Basic</i>	<i>Basic</i>
<i>Approaching Basic</i>	<i>Approaching Basic</i>
<i>Foundational</i>	<i>Unsatisfactory</i>
<i>Pre-Foundational</i>	

Thus, the upper 2 achievement level categories were designed to correspond with the *Approaching Basic* and *Basic* achievement levels from the LEAP/GEE tests. The area below the *Approaching Basic* cut (e.g., labeled *Unsatisfactory* on the LEAP/GEE tests) would be divided into two regions in order to better differentiate LAA 2 student performance in this region. It was hoped that this differentiation of students below the *Approaching Basic* level, along with the improved targeting of the LAA 2 test toward this student subpopulation, may provide better information for instructional purposes (including remediation).

The final recommended standards were converted into scaled score ranges. The scaled score range of the *Approaching Basic* achievement level was set to be exactly the same as the scaled score ranges found in the corresponding LEAP/GEE tests. The beginning scaled score for the *Basic* achievement level is also exactly the same as with LEAP/GEE. The top end of the *Basic* achievement level was truncated at a scaled score of 340 in all cases since the LAA 2 assessment was not designed to accurately assess students who may be emerging into the *Mastery* achievement level.

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

The primary purpose of form equating is to establish score equivalency between two (or more) forms. Equivalency is established by placing the forms 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.

Because spring 2006 represented the first operational administration of the LAA 2 tests there were no equating procedures that were necessary to maintain the underlying scale.

## ***Validity***

Criterion-referenced tests, such as the LEAP and GEE are designed to blueprints specifying the proportion, or weight, of the test in terms of score points that are devoted to any given content unit, such as a strand or standard. The blueprint, then, is used as a guide by test developers in assembling a test from a pool of candidate items that are classified by content unit. A test form is considered to be valid in terms of content if a form's score points for each content unit are similar to those specified by the test blueprint. Evidence of LEAP and GEE content validity can be found in the LEAP and GEE technical reports.

The LAA 2 test forms are configured slightly differently from the LEAP and GEE test forms. The test design of LAA 2 is a modification of LEAP and GEE's test design. The intent of the LAA 2 test was to maintain the essence of the original LEAP and GEE blueprints and content validity by including a proportion of items mandated by the LEAP and GEE blueprints and selecting the items from the same candidate pool.

## ***Reliability***

Reliability describes the accuracy of the test scores. There is some error associated with any test score. The more reliable the test, the less error is associated with that test score. The table below provides the scale score statistics, from the calibration sample, for the spring 2006 test administration. Reliability is reported in the last three columns of the table. The traditional method, Cronbach's alpha, is reported; however, given the assumptions of this method and the characteristics of the tests, this method typically underestimates the reliability of the test. Hence, a second form of reliability is computed, the Stratified alpha (Qualls, 1995). 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. The test reliabilities for LAA 2 are lower than the comparable LEAP or GEE tests, for the same grade and subject area, due to the shortened test lengths of the LAA 2 test configuration. The small LAA 2 test-taking population, in spring 2006, may also be a contributing factor to the lower reliabilities. Spearman-Brown prophecy formula allows us to calculate the estimated reliability of the LAA 2 tests if they were the same length as their LEAP or GEE counterparts.

Typically reliability estimates greater than .80 are considered very good, and above .85 excellent. It should be noted, however, that with shorter tests, smaller populations, and a subpopulation that may be more diverse and more homogeneous in their achievement level, that the somewhat lower reliability estimates of the LAA 2 tests are not surprising

### Number Correct Test-Level Summary Statistics

Grade Content	Form	Number of Items	Total Score Points	Mean P-Val	NC Mean	NC Standard Deviation	NC SEM	Reliability		
								Stratified	Cronbach	Spearman Brown Estimate
<b>4</b> <b>ELA</b>	1	26	35	.37	12.84	5.483	2.57	.781	.75	.85
<b>4</b> <b>MA</b>	1	44	46	.43	19.32	8.48	3.00	.875	.87	.91
<b>8</b> <b>ELA</b>	1	26	35	.41	14.47	5.48	2.66	.764	.74	.85
<b>8</b> <b>MA</b>	1	44	46	.36	16.14	6.17	3.04	.758	.75	.84
<b>10</b> <b>ELA</b>	1	26	35	.37	13.23	5.43	2.64	.764	.73	.84
<b>10</b> <b>MA</b>	1	44	46	.31	13.85	4.82	2.98	.618	.62	.73
<b>11</b> <b>SC</b>	1	37	39	.34	13.02	4.50	2.87	.593	.59	.68
<b>11</b> <b>SS</b>	1	34	36	.38	13.70	4.33	2.70	.612	.61	.77

### *References*

- Linacre, J. M. 2004. WINSTEPS [Computer Program]. Chicago: WINSTEPS.com
- Lineacre, J.M. and Wright, B. (2001). *A Users Guide to WINSTEPS*. Chicago: MESA Press.
- Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47, 149-174.
- Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research (Expanded Edition, 1980. Chicago: University of Chicago Press).
- Wright, B., and Masters, G. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
- Wright, B., and Stone, M. (1979). *Best Test Design*. Chicago, MESA Press.