



LEAP Alternative Assessment, Level 1 (LAA 1)

2013 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

In 1998, a state program called *Reaching for Results* began; its purpose was to raise achievement for all Louisiana students. While all Louisiana students are included in *Reaching for Results*, there is a small percentage of students with significant disabilities for whom the general statewide assessment is not appropriate. Therefore, an alternate assessment was created. The LEAP Alternate Assessment, Level 1 (LAA 1) has been specifically designed to evaluate the progress of students with significant cognitive disabilities.

The development process for the LAA 1 program began in 1998 when a group of general and special education educators and parents created the *General Educational Access Guide*. The LEAP Alternate Assessment (LAA) was administered for the first time in 2001. In 2006, LAA was renamed LEAP Alternate Assessment, Level 1 (LAA 1) with the inception of LEAP Alternate Assessment, Level 2 (LAA 2), a program designed for students with persistent academic disabilities. NCLB requires that assessment for students with significant cognitive disabilities be 1) academic-based, 2) aligned to content standards, and 3) at grade level or grade spans. In October 2006, the Louisiana Department of Education (LDE) was advised by the United States Department of Education peer review to redesign LAA 1 to meet these requirements and to implement the program by spring 2008. The LAA 1 program was redesigned and developed in 2007 and administered for the first time in the spring of 2008.

LAA 1 is designed to evaluate students whose Individualized Education Programs (IEPs), reflect significant modifications of the general education curriculum, yet emphasize academic standards as well as functional and life skills. LAA 1 is a performance-based assessment that evaluates each student's knowledge and skills on the Extended Louisiana Content Standards. It is a standardized assessment in that each student in the same grade span is administered the same tasks. The test administrator uses item-specific rubrics to score the student's performance. LAA 1 is assessed in three subject areas: English language arts, mathematics, and science. A student

participating in LAA 1 is progressing toward a Certificate of Achievement rather than a state high school diploma.

A number of groups and organizations are involved in the development of the LAA 1 program, including Louisiana Department of Education (LDE), Louisiana educators and communities, Louisiana Technical Advisory Committee (TAC), and Data recognition corporation (DRC) as the contractor. Each of the major contributors serves a specific function, and their collaborative efforts contribute significantly to the testing program's success.

This technical summary provides brief information on the LAA 1 development and evidence of technical quality of the LAA 1 program from the 2013 spring operational administration. These tests were administered to all eligible students in Louisiana public schools in English language arts and mathematics at grade spans 3–4, 5–6, 7–8, and grade 10, as well as in science at grades 4, 8 and 11 in February and March 2013.

Extended Standard

The Louisiana content standards and the grade level expectations (GLEs) were developed for use by regular education students and most students with disabilities. To meet the needs of students with significant cognitive disabilities, the LAA 1 Extended Standards (ESs) were developed in July-August 2007.

The development of the Extended Standards began by identifying the core academic content considered appropriate for students taking LAA 1 for each grade span and subject area. LDE with the assistance of special education consultants, made an initial selection of the appropriate standards and benchmarks to be extended. DRC then prepared draft Extended Standards aligned to these benchmarks. Once appropriate benchmarks and GLEs were identified, the core academic content was extracted and the appropriate cognitive demands (expectations) were identified.

Each Extended Standard provides a description of the essence of a content standard and the GLEs appropriate for students who meet the eligibility criteria for LAA 1. Additionally, three levels of academic complexity related to each ES provide instructional access for students with varying academic abilities. Extended Standards have been developed for English language arts and mathematics in grade spans 3–4, 5–6, 7–8 and grade 10 as well as for science grades 4, 8, and 11.

Test Development

The ESs provide a way for students with significant cognitive disabilities to access the general education curriculum, and the foundation for the redesigned LAA 1 program. Under the guidelines of ESs, series of important steps proceeded through in the development and administration cycle in 2007-08 to make sure LAA 1 valid, reliable and unbiased, including (1) test design, (2) development of test blueprint, (3) development of performance tasks, (4) Louisiana educators' reviews, (5) revision, (6) operational form construction, (7) alignment of test content to the extended standards, (8) operational administration, and (9) standard-setting for LAA 1 achievement levels, and (10) operational test data analysis.

Test Design and Blueprints

LAA 1 is composed of all performance tasks. These performance tasks are graphic by design and each is accompanied by a script that is read to the student by the administrator. Each performance task is scored on a 0-1 point or a 0-2 point scale, according to an item-specific rubric. The test design is limited to the use of 25 performance tasks per subject and grade span to reduce the impact of assessment administration on this population.

LAA 1 test blueprints and reporting categories are similar to those for the standard Louisiana academic assessments, such as LEAP/GEE and *i*LEAP. In addition, five possible score points were established by LDE as the minimum points necessary per reporting category to provide for reasonably reliable reporting of test results. Because the number of performance tasks per grade span and subject was limited to 25 for pedagogical reasons, a minimum of 35 points per subject and grade span was established as a target for test blueprint development. The LAA 1 test blueprints are shown in Appendix B of the 2013 LAA 1 Technical Report.

Performance Tasks (Items) Development

Following the Extended Standards committee meeting (August 27–29, 2007), an item development plan was developed to ensure that sufficient items reflecting the depth and breadth of the Extended Standards and complexity levels were developed for a range of estimated difficulty levels. Given the communications modalities of students with significant cognitive disabilities, a script-based spoken delivery model with graphics or limited text supported by graphics was the selected item format for LAA 1. In addition, two test booklets, one for the test administrator and one for the student, were the selected administration format.

Performance tasks were developed by a team of DRC special education and test development specialists following the Item Development Guidelines for Alternate Assessment. As sets of items were finalized they were reviewed by LDE special education staff, which provided DRC item writers with ongoing feedback used to refine the performance tasks prior to the committee review. During this process, LDE feedback was proactively applied to new performance tasks as they were developed.

Before the drafts of performance tasks were reviewed by LDE and Louisiana educators, DRC content specialists, special education staff, and test development editors conducted an internal item review process to ensure that sufficient tasks meet the test blueprint requirements for each subject and grade-span; that all tasks are aligned to the content and cognitive requirements of the applicable Extended Standard and complexity level; that all tasks are free of errors and typos, bias, and sensitivity concerns; and that all tasks for a subject and grade-span provide for a range of content, cognitive complexity, and difficulty.

A content/bias and sensitivity review committee representative of the Louisiana K–12 education community (general and special education teachers) by grade level, geographic distribution, gender, and ethnicity, LDE assessment professionals, and DRC Test Development staff reviewed the draft performance tasks. The committees reviewed 353 performance tasks and related passages and accepted as modified 342 performance tasks, a 97% survival rate.

Test Form Construction

During the face-to-face meeting with LDE following the content review meeting with Louisiana educators, 25 performance tasks per subject and grade span were selected for inclusion in the operational form. The following provides the general guidelines used to construct LAA 1 test forms.

- The distribution of item content, number of items, and total points possible shall be consistent with the requirements of the test blueprints for each subject and grade-span.
- Item selection should provide for a range of estimated difficulty consistent with the academic achievement levels.
- Items selected should not be repetitive in nature and should address the broadest range of content and cognitive complexity.
- Items should be free of clang and cluing between items should be minimized.
- Items should be sequenced by a combination of estimated difficulty and content themes allowing for a natural progression of content and difficulty based on professional judgments.
- Final scripts, item instructions, and scoring rubrics should be consistent, use common and standardized language, and be similar in format to facilitate ease of administration.
- Adequate linkage should be provided between task scripts in the Administrators Booklet and the tasks in the Student Booklet to insure that students are attempting the correct task when the task is read aloud.
- All unnecessary text and/or graphics should be deleted from the test booklets.

Test Content Alignment

After LAA 1 test forms are completed, a test content alignment study was conducted for English language arts and mathematics across grade spans 3-4, 5-6, 7-8, and 9-10 and for science across grades 4, 8, and 11 in January 2008. Special education expert reviewers from Louisiana and national expert reviewers participated in this study. The final results of the alignment study indicated that there was a strong alignment between the LAA 1 Extended Standards and the corresponding LAA 1 tests across subjects and grade spans (or grades).

Test Administration

Participation Criteria

The decision to test a student with LAA 1 is not based on the student's placement, or solely based on the student's disability according to Bulletin 1508, or excessive or extended absences, or social, cultural, and/or economic differences. Nor is this decision based on its participation impact on school performance scores. This decision is an IEP team decision based on the needs

of the student; it is not an administrative decision. The following criteria were used to determine whether a student may participate in LAA 1:

- The student’s impairments cause dependence on others for most, if not all, daily-living needs and the student is expected to require extensive, ongoing support in adulthood.
- The student’s instructional program emphasizes life skills and functional applications of the general curriculum.
- The student requires extensive instruction on functional skills in multiple settings (e.g., school, work, home, community) to acquire, maintain, and generalize skills necessary for application in school, work, home, and community environments.
- Current longitudinal data (e.g., classroom observation, task analyses, progress on IEP objectives, evaluations, and parental information) indicate the student should participate in LAA 1.

Test Administration Guides and Training

A *LAA 1 Test Administrator Manual* describes assessment administration procedural details. Task-specific scripts for administering each Performance Task were included in the Administrator Booklet. In addition, LDE provided several professional development workshops for LAA 1 administrators prior to test administration to explain the reasons for the redesign of the LAA 1 assessment, the Extended Standards, the new LAA 1 assessment design, procedures for administering the assessment, the testing window, test accommodations for students with special needs based on IEP provisions, recording of student responses, and other matters pertaining to the revised scope of the LAA 1 assessment program.

Test Accommodations

The test accommodations for students taking other statewide assessments, such as large printing, tests read aloud, communication assistance, extended/adjusted test-taking time, answer-recorded, and individual/small group testing, have been incorporated into the design of LAA 1. Most of the accommodations for the LAA 1 population are related to the use of *Assistive Technology*.

Two main requirements for the use of accommodations are that test accommodations must not be different from or in addition to the accommodations documented on the student’s IEP and provided in regular classroom instruction and assessment and that test accommodations must not breach test security or invalidate the meaning of the test score or the purpose of each performance task.

Test Reporting

The following types of scores were provided for the LAA 1 reporting:

Raw Score A raw score is the number of points a student earned in a subject-area test. By itself, the raw score has limited utility; it can only be interpreted with reference to the total number of items on a subject-area test.

Scaled Score Scaled scores are statistical conversions of raw scores that adjust for slight shifts in item difficulties and permit valid comparisons across all test administrations within a particular grade and subject. Students in grade 5 who achieved a scaled score of 850 on the 2012 LAA 1 ELA test demonstrated the same level of knowledge and skill as students in grade 5 who achieve a scaled score of 850 in 2013. With scaled scores, schools can compare the demonstrated knowledge and ability of different groups of students and/or across years. Comparing scaled scores on LAA 1 can help schools determine the impact of instruction and curriculum.

Performance Levels Three achievement levels were incorporated into the LAA 1 assessment program: *Exceeds Standard*, *Meets Standard*, and *Working Toward Standard*. The general definitions of the achievement levels are as follows:

- **Exceeds Standard:** A student at this level has demonstrated *expanded* academic knowledge and skills included in the grade-level Extended Standards.
- **Meets Standard:** A student at this level has demonstrated *fundamental* academic knowledge and skills included in the grade-level Extended Standards.
- **Working Toward Standard:** A student at this level has demonstrated *minimal or inconsistent* academic knowledge and skills included in the grade-level Extended Standards. However, the student may be developing introductory academic knowledge and skills that can be built upon to access the grade-level curriculum.

The scaled score cut for *Meets Standard* across grade spans and subjects is 810 on the 700–900 scale, while the cuts for *Exceeds Standards* vary across subjects and grade spans.

Content Standard Score (Strand Score) A content standard score describes a student's or school/district's performance on a particular content standard defined in the test. In LAA 1, content standard scores are derived from percent correct, indicating the percentage a student or a school/district completed correctly in a test. Content standard scores are helpful in identifying a student's or a group's strengths and weaknesses on the test. Also, it is useful to compare average content standard scores of a school against a norm or reference group (e.g., the state average or system average).

Louisiana schools and districts received the following reports in ELA, mathematics and science: Online Student Report, Student Label, School Roster Report, Achievement Level Report, Special Education Exceptionality Report, Subgroup Report, and Interpretive Guide.

Performance Standards

The primary goal of the standard-setting meeting was to use an acceptable and defensible standard-setting methodology to recommend the level of knowledge, skills, and ability required to identify an achievement level of each student. A key component of this methodology was the utilization of Alternate Achievement Level Descriptors (AALDs) to delineate the LAA 1 achievement levels based primarily on the academic criteria described in the Extended Standards.

The AALDs are content and grade-level specific criteria, based on the Extended Standards that describe what Louisiana students taking LAA 1 should know and be able to do at each

achievement level. The AALDs were used extensively by the panelists during the standard-setting process to identify where cut scores should be placed from a content perspective.

The bookmark procedure was used for LAA 1 standard-setting. The Bookmark method has been used successfully in Louisiana for the establishment of LEAP/GEE, iLEAP and LAA 2 achievement levels. In this procedure, panelists are presented with items ordered from easiest to most difficult and are asked to place a bookmark to make cut score judgments directly onto the LAA 1 score scale, in the context of item content, the Extended Standards, and the LAA 1 AALDs. The cut scores separate content which should be mastered from that which is not necessary to master for a given achievement level. Following several rounds of consideration, final cut scores were established by determining the median value of cut scores across the individual panelists.

The bookmark standard-setting process was set up in a series of rounds. All panelists' standards recommendations were made as individuals, but the rounds included discussions of the AALDs at both a small-group level and room level. Information was provided on the percentages of students in each achievement level based on the median of the panelists' most recent recommendations. It was anticipated that all groups for LAA 1 would complete three rounds of the process.

After standard-setting at all grade spans and subjects, an Articulation Committee, consisting of table leaders from each of the standard-setting subcommittees and contractor staff, convened to examine the cut scores across grade-span levels and content areas for reasonableness. A review of the final panelists' recommended standards was done by psychometric staff and some minor adjustments were made to a few of the panel's recommendations. These adjustments were within small statistical margins of error afforded by the standard-setting process.

The LAA 1 standards advisory meeting was held on July 8–10 2008 in Baton Rouge, Louisiana. The purpose of this meeting was to have educators make recommendations on standards for LAA 1 following the bookmark method and using their classroom expertise. The LAA 1 standards advisory panelists consisted primarily of educators and/or district personnel with a mixture of expertise in both special education and general education. To ensure that all panelists had the same instructions about the process, all panelists received training as a large group on the morning of the first day of the meetings. Detailed information of LAA 1 standard-setting is included in the 2008 LAA 1 Standard-Setting Technical Report.

Measurement Model

Given that LAA 1 scores are based on performance rating scales (0–1 or 0–2 point scale) rated by teachers and there is no guessing factor, the Rasch Partial Credit Model (RPCM) (Wright and Masters, 1982) may be the best fit with the LAA 1 assessments. The RPCM model is an extension of the Rasch, one-parameter IRT model attributed to Georg Rasch (1960), as extended by Wright and Masters (1982).

The RPCM was selected because of its flexibility in accommodating both score scales (0–1 and 0–2) 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 years. The RPCM is defined via the following mathematical measurement model where, for a given task involving m score categories, the probability of person n scoring x on a polytomous items or prompt i is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^x (B_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (B_n - D_{ij})}, \quad (1)$$

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

$$\sum_{j=0}^K (B_n - D_{ij}) = 0. \quad (2)$$

The RPCM provides the probability of a student scoring x on the m_i step of item or 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 item or prompt i (Wright and Masters, 1982). Note that for tasks on a rating scale of 0–1 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 a task difficulty.

The application of the RPCM permits all of the tasks either with 0–1 score points or with 0–2 score points to be placed on the same scale. This means that all common task- and step-difficulty estimates will be on the same underlying logistic scale as that of the student proficiency level estimates.

At the conclusion of these calibrations, all task difficulty estimates, as well as all student proficiency level estimates, are directly comparable on the same underlying logistic scale. Therefore, under the RPCM scaling, all tasks, independent of scoring type, are placed onto the same common score scale.

The scaled score system used for LAA 1 was derived from Rasch abilities obtained from the baseline test form. The baseline test form is the 2008 spring administration in the specific content area. The purpose of a scaled score system is to convey consistent information about student performance from year to year and across administrations within a year. Because test items composing different test forms will be of different difficulty, a statistical “equating” process will be used for the future new forms to ensure the scores are comparable.

The scale for LAA 1 ranges from 700 to 900 for all content areas tested from the lowest of 700 to the highest of 900. Using the scaled score ensures that the amount of knowledge required to achieve the performance level *Meets Standard* remains the same, regardless of when a student took the test.

Validity

Validity is the process of collecting evidence to support inferences from the use of the resulting scores from an assessment. Assessment results must show evidence of validity. For LAA 1, the score use is applied to knowledge and understanding of the Louisiana content standards and the Extended Standards (ESs).

Content/Curricular Validity

Content/curricular validity requires evidence based on test content. As a criterion-referenced assessment based on an extensive definition of the content it assesses; LAA 1 is content based and aligned directly to the statewide content standards. Content validity can be addressed in the followings:

Relation to the State Content Standards

From the inception of LAA 1, a committee of educators, item development experts, assessment experts, LDE content specialists and psychometricians, and DRC content specialists and psychometricians worked together in the following steps: (1) identified the core Louisiana academic content standards for LAA 1 population, (2) extended the identified standards for this population and developed the Extended Standards (ESs), (3) developed the LAA 1 test design and blueprints under the guidelines of BSs, (4) developed the item development plan to ensure sufficient numbers of items reflect the depth and breath of ESs, and (5) reviewed all performance tasks. All of these activities provide multiple opportunities to ensure LAA 1 items and tests measure appropriate content and offer insights on the interpretation of the statewide content standards. In addition, independent alignment studies were conducted to examine the alignment between the items and the content standards they were intended to measure. The nature and specificity of these activities and procedures provided strong evidence for the content validity of LAA 1.

Item-to-Content Area Match

Expert judgments from educators, test developers, and assessment specialists supported the alignment of LAA 1 with the state Extended Standards. In addition, because expert teachers in the content areas were involved in establishing the content standards, the judgments of these same expert teachers in the review process provided a measure of content validity. A match between the content standards and the components of the LAA 1 proves that the assessment measured the content standards. A table showing the number of assessment components, tasks, or items matching each content standard is conventionally used to document evidence for the content validity of an assessment. The LAA 1 test blueprints, as shown in Tables 10.1.A through 10.1.C in the 2013 LAA 1 Technical Report, provide such documentation.

Construct Validity

Construct validity refers to the degree to which the test score is a measure of the psychological characteristic (i.e., construct) of interest. A construct is an individual characteristic assumed to exist in order to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular

characteristic from the assessment results is inferred, a generalization or interpretation of some construct is made. Evidence of construct validity is collected as follows:

Standards Intercorrelations

The LAA 1 content domains are subdivided into smaller content units, referred to as standards, strands, or reporting categories. Intercorrelations among the standards provide evidence of construct validity and/or convergent validity. Test blueprint specifies the number and type of items associated with each standard for purposes of test construction. Most intercorrelations are in the expected range, for standards with more than four score points.

Item-Total Correlation (or Point-Biserial Correlation)

Metrics of construct validity for the LAA 1 are item-total correlations. An item-total correlation is the correlation between an item and the total test score. Conceptually, if an item has a high item-total correlation, it indicates that students who performed well on the test overall got the item right and students who performed poorly on the test overall got the item wrong. Presuming that the total test score represents the extent to which a student possesses knowledge of the construct being measured by the test, high point-biserial correlations indicate that the items on the test require knowledge of this construct in order to be answered correctly. Results of item-total correlations show strong evidence of construct validity for LAA 1 forms.

Validity Evidence for Different Student Populations

Because LAA 1 assesses the state standards and Extended Standards required to be taught to all students, the tests cannot be more or less valid for use with one subpopulation of students over another subpopulation. In other words, because the LAA 1 measures what is required to be taught to all students and was administered under the same standardized conditions to all students, the validity of score interpretations should apply to all students.

Great care has been taken to ensure LAA 1 items are fair and unbiased for all students. Much scrutiny was applied to the items and their possible impact on minority or other sub-populations making up the population of the state of Louisiana. Every effort was made to eliminate items that may have ethnic or cultural biases.

Evidence supporting the fairness of LAA 1 can be found from the analyses that examine the relationships between student scores and student demographic characteristics. Also, the scaled score means and standard deviations of student subpopulations, such as gender, ethnicity, lunch status, limited English proficiency, and special education classifications, were analyzed and compared. The results do not show the clear patterns that are commonly observed with regular education students.

Reliability

Reliability describes the accuracy of the test scores. The more reliable the test, the less measurement error is associated with that test score. The table on the next page provides the number correct score statistics, from the research sample, for the spring 2013 test administration. 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, 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.

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 approximately 68% of the time a student's true score would fall within one SEM interval so constructed.

| Content Area | Grade Span/ Grade | # of Items | Total Score Points | Form Mean | Form Standard Deviation | Form SEM | Stratified | Cronbach |
|---------------------|--------------------------|-------------------|---------------------------|------------------|--------------------------------|-----------------|-------------------|-----------------|
| ELA | 3-4 | 25 | 36 | 19.50 | 9.64 | 2.51 | 0.93 | 0.93 |
| | 5-6 | 25 | 36 | 20.84 | 9.76 | 2.49 | 0.94 | 0.94 |
| | 7-8 | 25 | 36 | 21.74 | 10.15 | 2.51 | 0.94 | 0.94 |
| | 10 | 25 | 35 | 21.47 | 10.10 | 2.49 | 0.94 | 0.94 |
| Math | 3-4 | 25 | 35 | 20.30 | 10.47 | 2.52 | 0.94 | 0.94 |
| | 5-6 | 25 | 35 | 20.93 | 9.81 | 2.46 | 0.94 | 0.94 |
| | 7-8 | 25 | 35 | 21.79 | 9.98 | 2.34 | 0.96 | 0.95 |
| | 10 | 25 | 35 | 20.68 | 9.96 | 2.44 | 0.95 | 0.94 |
| Science | Grade 4 | 25 | 43 | 24.47 | 11.77 | 2.74 | 0.95 | 0.95 |
| | Grade 8 | 25 | 42 | 26.58 | 11.30 | 2.50 | 0.95 | 0.95 |
| | Grade 11 | 25 | 48 | 29.10 | 13.31 | 2.70 | 0.96 | 0.96 |

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 66, 443-459.
- Brauen, M. L., O'Reilly, F., & Moore, J. (1994). *Issues and options in outcomes-based accountability for students with disabilities*. Rockford, MD: Westat, Inc.
- Brennan, R. L. (2001a). *Manual for mGENOVA*. Iowa City: Iowa Testing Programs, University of Iowa.
- Brennan, R. L. (2001b). *Generalizability theory*. New York: Springer Verlag.
- Cronbach, L. J., Linn, R. L., Brennan, R.L. & Haertel, E. H. (1997). Generalizability analysis for performance assessment of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Elliot, S. N. & Roach, A. T. (2007). Alternate assessments of students with significant disabilities: Alternative approaches, common technical challenges. *Applied Measurement in Education*, 20(3), 301-333.
- Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting Performance Standards*, (pp. 219-248). Mahwah, NJ: Lawrence Erlbaum Associates.
- Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15-40.
- Lee, W. C. (2008a). *Classification Consistency and Accuracy Under the Compound Multinomial Model*. (CASMA Research Report No. 13). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from <http://www.education.uiowa.edu/casma>).
- Lee, W. (2008b). *Program MULT-CLASS (Version 3.0) for Multinomial and Compound-Multinomial Classification Consistency and Accuracy*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from <http://www.education.uiowa.edu/casma>).
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26, 412-432.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard-setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the meeting of the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

- Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessing in teaching* (7th ed.). New Jersey: Prentice-Hall Inc.
- Linacre, J. M. (2004). *WINSTEPS, Rasch-Model Computer Program*. Chicago: MESA Press.
- Linacre, J. M. (2004). *A User's Guide to WINSTEPS MINISTEP, Rasch-Model Computer Programs*. Chicago: MESA Press. <http://www.winsteps.com>.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational measurement*, 32, 179-197.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Louisiana Department of Education. (2007). *2006–2007 Annual Report, LEAP Alternate Assessment, Level 1*. Louisiana Department of Education.
- Louisiana Department of Education. (2008). *2007–2008 Annual Report, LEAP Alternate Assessment, Level 1*. Louisiana Department of Education.
- Louisiana Department of Education. (2009). *2008–2009 Annual Report, LEAP Alternate Assessment, Level 1*. Louisiana Department of Education.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., and Bock, R. D. (1994). *PARSCALE Version 2.0*. Mooresville, IN: Scientific Software.
- Perie, M. (2007). *Setting alternate achievement standards*. Lexington, KY: University of Kentucky, National Alternate Assessment Center. Retrieved February 10, 2008 from <http://www.naacpartners.org/products.aspx>
- Qualls, A. L. (1995). Estimating the Reliability of a Test Containing Multiple Item Formats. *Applied Measurement in Education*, 8, 111-120.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, 1980, Chicago: University of Chicago Press).
- Sireci, S. G., Hambleton, R. K. & Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review*, 15(1), 21–25.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175-186.
- U.S. Department of Education. (April 2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author.

U.S. Department of Education. (December 2007). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Revised December 21, 2007, to include modified academic achievement standards. Washington, DC: Author.

Wright, B. D. and Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.