Listening. Learning. Leading.®

Research Report
ETS RR–13-20

Evaluation of a Condition-Adaptive Test of Reading Comprehension for Students With Reading-Based Learning Disabilities

Elizabeth Stone

Linda Cook

Cara Laitusis

October 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Evaluation of a Condition-Adaptive Test of Reading Comprehension for Students With Reading-Based Learning Disabilities

Elizabeth Stone, Linda Cook, and Cara Laitusis
ETS, Princeton, New Jersey

October 2013

Action Editor: Donald E. Powers

Reviewers: Yigal Attali and Eric Hansen

Abstract

This study presents secondary analyses on a 2-stage test of reading comprehension for students with reading-based learning disabilities (RLD). The present paper describes student perceptions of the test and its features as well as analyses focused on the routing test and associated cut score. The routing test contained typical state assessment content and was designed to route the RLD participants into 1 of 2 accessible 2nd-stage tests, 1 accommodated and 1 not accommodated, and we refer to this structure as *condition-adaptive*. The accommodated test was presented with a read-aloud accommodation, and an oral reading fluency subtest was also administered as part of that test to evaluate decoding skill that may have been masked by the read-aloud accommodation. This design allowed for componentwise measurement of reading proficiency at the 2nd stage with the goal of creating an accessible tailored test that would provide more comparable accommodated and nonaccommodated test scores for possible use for accountability. Overall, student perceptions differed depending on which 2nd-stage test was taken, but most students indicated that they tried as hard on this field test as they did on other tests, and many expressed a preference for the read-aloud accommodation. We found that routing test length was more strongly correlated with 2nd-stage test performance for the higher performing RLD students, that relatively similar routing decisions could be made using a shorter routing test made up of a smaller subset of the available passages, and that the percentage of students routed into each 2nd-stage test was sensitive to perturbations of the cut score.

Key words: accessibility, condition-adaptive test, multistage test, read-aloud accommodation, routing accuracy, routing test

Acknowledgments

Funding was provided by Grant No. H324F040001 from the U.S. Department of Education's Office of Special Education Programs (OSEP). Portions of this paper were presented at the National Council on Measurement in Education meeting, April 2011, in New Orleans, LA.

Table of Contents

List of Tables

Page

List of Figures

Page

Accountability testing in the K-12 sector has as one goal "[t]o close the achievement gap with accountability, flexibility, and choice, so that no child is left behind" (NCLB, 2001). NCLB legislation requires states to evaluate student proficiency via approved assessments aligned with content-based standards. This accountability framework has been difficult to implement equitably for all students and is in the process of undergoing reform (Sunderman, 2006). To address this issue, some individual states now use adaptive testing, and one state consortium has begun the development of an adaptive testing framework to meet accountability needs (REL West/WestEd, 2008).

Test delivery mechanisms fall into three basic categories: *linear*, *item-level adaptive*, or *multistage*. Most state tests are of the first type, linear, in that the same scored items are given to all test takers in the same order (unscored items—pretest items, for example—may vary between test takers). On the other end of the spectrum are item-level adaptive tests. In an item-level adaptive test, the item selection mechanism can take into account estimated ability after each item response and can then select the next item based partially on that estimate, allowing the algorithm to fine-tune the estimate of ability (i.e., where the test taker falls on the proficiency spectrum). Adaptive testing provides a possible solution, in theory, to the test difficulty issues discussed previously. Some states already use, or are planning to use, adaptive tests or adaptive portions of tests to measure student proficiency. The third type of test, a hybrid between fully linear or item-level adaptive tests, is referred to as multistage. The term *multistage* has been used broadly and encompasses various test designs that incorporate more than one test stage with some dependence between stages. A multistage test adapts the difficulty of the assessment using a series of linearly administered subtests. In a multistage test, the ability estimate after one section can be used to route the student into a more or less difficult section, better targeting the difficulty of the majority of the test. Multistage testing forms a compromise between linear tests and item-level adaptive tests, with some of the best features of each.

Tests with adaptive elements provide better targeting of the test to individuals. This may be of particular benefit for many students with disabilities, for whom it has been suggested that state tests do not appropriately measure proficiency. First, although there has been much improvement in this area, state tests are not always designed following principles of universal design (Johnstone, Altman, & Thurlow, 2006) that incorporate accessibility into test development from the initial stages. Students may not find a test accessible even if they have

some proficiency in the overall subject matter. Obstacles to accessibility for students with disabilities may involve direct physical barriers such as the inability to see test content and no alternative way to engage with that content. For students with reading-based learning disabilities (RLD), the inability to decode text could present an accessibility barrier to demonstrating proficiency in reading comprehension. However, depending on the operational definition of reading, providing an audio accommodation that eliminates the need for the foundational or gateway skill of decoding in the process of comprehension may invalidate the resulting test score (Johnstone & Thurlow, 2010). Consequently, efforts to make tests more accessible to students with disabilities often involve changes to the test or test administration (e.g., a read-aloud or audio accommodation) that may affect construct measurement and test score interpretation. As of 2011, 37 state assessment policies prohibited the use of read aloud on reading assessments, and 15 state assessment policies allowed its use but did not include those scores when reporting adequate yearly progress measures (Thurlow & Larson, 2011).

Second, state accountability tests must measure, in a reasonable number of items, knowledge of constructs covering a broad curriculum with specific standards. In most cases, the testing population includes all students except those with severe cognitive impairments, who are not expected to perform at grade level; thus, the test must assess students with a wide range of ability, and students with learning disabilities may fall into the lower tail of the proficiency distribution with scores around chance level (Minnema, Thurlow, Bielinski, & Scott, 2000). A multiple-choice (MC) test of 100 four-option items for which students have an average score of 25 is a simple example of this phenomenon. Essentially, student scores cannot be distinguished from those that would be obtained were the students to guess on the test items. An assessment for which this is true provides inadequate information about proficiency for those students. In addition, the difficulty of the test may induce anxiety in and decrease the engagement of students who are unable to demonstrate proficiency, which may compound the measurement problem. For all of these reasons, it is worthwhile to pursue the goal of developing tests that are accessible to students with RLD and that produce comparable scores for all students. One possible avenue in that direction is the use of more flexible modes of test delivery.

Tests with adaptive elements may be more efficient, requiring fewer items to determine student proficiency reliably, shortening testing time, and often allowing for more rapid score reporting (Goldstein, 2003; Thompson & Way, 2007; Trotter, 2003). Reduced testing time may

prevent the onset of fatigue, a benefit to all students. However, while adaptive testing may also be better able to target test difficulty for students who would normally be ill-measured by a linear state test, there is some concern that moving to easier items to assess the students with poorest performance may lead to an out-of-level test (Kingsbury & Hauser, 2004; Thompson & Way, 2007; Thurlow, Elliott, & Ysseldyke, 2003). The inclusion of items that reflect below-grade-level content conflicts with the NCLB requirement that accountability test scores reflect knowledge of on-grade-level content. However, some proposals for the reauthorization would allow for flexibility in this area.

Another option for increasing accessibility is to introduce flexibility in whether a test is administered with a read-aloud (or other) accommodation, adapting the test condition based on a measure of whether the test taker needs it for that particular construct and type of test content. By introducing a complementary proxy measure of any skills that are masked by the accommodation (e.g., decoding, for a read-aloud accommodation), the scores between accommodated and nonaccommodated conditions will be more comparable. By isolating the distinct components of reading (e.g., comprehension or decoding) into subtests, an accommodation can be presented during individual subtests without causing validity issues for all parts contributing to the test score. Further componentwise scores may be reported, providing diagnostic feedback for instruction and an overall measure of reading proficiency. The condition-adaptive approach described here, an attempt to develop a prototype assessment that could be used to test students with RLD accessibly in an accountability context, was employed when building the Designing Accessible Reading Assessments (DARA) project (http://www.ets.org/research/dara/overview/) field test that was administered in early 2010.

The DARA assessment, which was delivered on paper, incorporated a nontraditional adaptive element by including a first-stage reading comprehension routing test that determined which of two second-stage reading comprehension tests RLD students took: accommodated with read-aloud, or not accommodated. The accommodated second-stage test also included an assessment of oral reading fluency (ORF) that provided a complementary measure of the decoding skill masked by the read-aloud accommodation. The first-stage test (criterion) was designed to approximate a typical state accountability test and provide a baseline measure. The second-stage tests are referred to as component tests because of the isolated measurement of decoding (through the ORF measure) and reading comprehension (through the passage-based

MC items). The routing procedure, based on student proficiency on construct-relevant material, was implemented to more appropriately assign portions of the RLD group to the accommodated test while providing a more accessible test to all students. See Figure 1 for a schematic describing the stages of the field test.

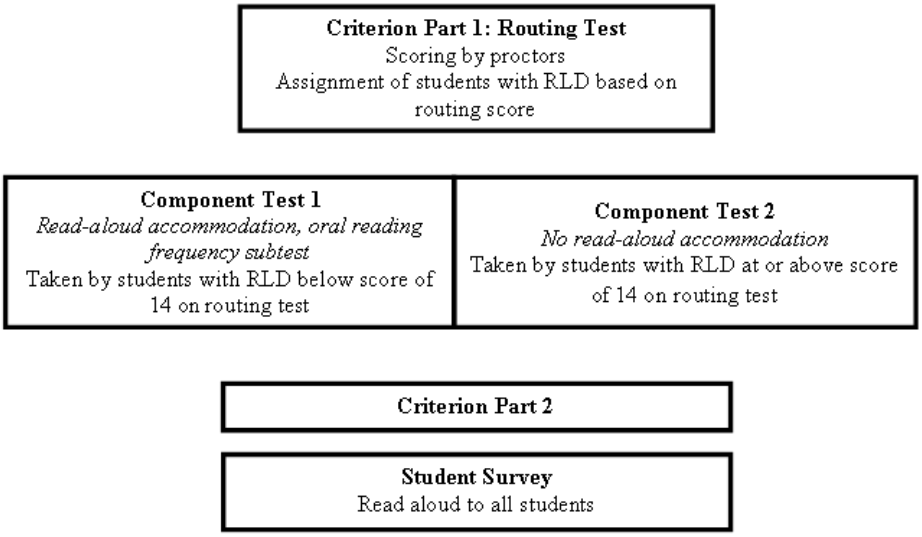

***Figure 1*. Designing Accessible Reading Assessments (DARA) field test design. RLD = reading learning disabilities.**

The DARA assessment field test was designed to compare the measurement properties of the component tests versus the criterion test for students with RLD and students without disabilities. We also wanted to determine whether there was a differential boost in scores between the two groups when examining gains between the criterion and component tests. Such a result would provide evidence to support the hypothesis that the read-aloud accommodation leveled the playing field by removing barriers to access rather than by reducing the difficulty of the test. Further, we wanted to determine whether the same construct was being measured by the two component tests. If so, that would be evidence to support the comparability of test scores and their aggregation for accountability purposes. This would be an advantage over the use of standard and accommodated tests that yield scores that cannot be combined. See Laitusis, Stone, Steinberg, and Cook (in press) for comprehensive results of the primary research questions.

While the design is loosely multistage, it differs from the traditional use of the term because the difficulty of the items was not explicitly varied between the two second-stage tests

and the scores from the stages were not combined. The second-stage test items in each of the two component tests were the same items but were delivered under different conditions (read-aloud or no read-aloud). This test versioning results in a potential difference in perceived and observed difficulty due to the accommodation, one reason that scores under the two conditions are not typically considered to be comparable on current state assessments. Further, the accommodated component test's ORF subtest contributed to the Component Test 1 score, making the second-stage tests different due to the addition of that subtest. See Stone and Bruce (2010) for more detail about the ORF subtest and student performance on it.

Relevant Research

Although multistage testing is not new (see, e.g., Cronbach & Gleser, 1965; Lord, 1971a; Lord, 1971b), its popularity has risen recently due to its use in well-known testing programs (e.g., the *GRE*® revised General Test). Multistage tests (MSTs) typically consist of a first-stage routing test and subsequent stages involving branches of tests at different difficulty levels (see, e.g., Hendrickson, 2007). An MST score may then be formed as a function of the aggregated scored responses at each stage, and different examinees may take different paths through the test. The items in each stage comprise structures that are often referred to as *testlets*, and the definitions of these structures vary (Hendrickson, 2007). Testlets may have balanced or homogeneous content and may or may not have stimulus-based or other dependencies. We will apply the term testlet loosely to denote a group of items administered together that constitute one stage of an MST for a test taker.

The multistage setup is partially adaptive, in that there are adapting points between each pair of stages that route test takers to the appropriate testlet at the next stage based on the current proficiency estimate. However, testlets are often designed to resemble a conventional test within each stage. In this way, multistage testing represents a range within the continuum from nonadaptive (i.e., conventional linear) to fully adaptive (e.g., computerized adaptive testing). Linear tests offer full control of content and the ability to ensure content balance, but the tests are usually targeted toward the middle of the proficiency distribution, which leads to poor measurement at the tails. Adaptive testing can more appropriately match item difficulty to estimated examinee proficiency; however, item selection and administration are optimized according to many different constraints in practice (e.g., content, item exposure, item overlap).

Adaptive tests have also been shown to increase student engagement by more appropriately targeting the overall test to an individual's ability range (Betz & Weiss, 1976).

MSTs capitalize on many of the advantages offered by linear and adaptive testing. They provide more control over content and difficulty level than do adaptive tests; consequently, MSTs may provide better measurement for groups at specific ability levels than do linear tests. Additionally, MSTs usually allow for students to review or change their item responses within testlet, while computer-assisted tests typically do not because each item response drives selection of the next item. MSTs can also reduce overexposure of the most informative items, which are typically those most often selected when an adaptive procedure involving maximum information is used (see, e.g., Eggen, 2001). However, MSTs may not provide the same accuracy in ability estimation or the same efficiency as do item-level adaptive tests (Patsula, 1999). Lord (1974) considered a multilevel *SAT*®, in which a routing test routed test takers to one of several tests with varying difficulty levels. In that illustration, the score on the routing test was not combined with the score from the tailored test because the routing test was self-administered and self-scored and was, therefore, not administered under comparable testing conditions. Most recent references consider adaptive MSTs that are based on item response theory and do provide a final score that is a combined score based on the multiple stages of the test. The literature on MST has focused on the number of stages, number of levels or testlets per stage, and number of items in the routing test or per testlet (Jodoin, Zenisky, & Hambleton, 2002; Weissman, Belov, & Armstrong, 2007; Zenisky, 2004). In general, there is evidence that a test composed of few stages and few testlets per stage will provide an increase in measurement quality over a linear test, as long as the routing test is sufficiently able to route test takers accurately and there is proper content coverage and difficulty range in subsequent testlets along each possible path. It should also be clear that the designs explored that fall into the broad category of MSTs are varied.

The topic of advantages and disadvantages of testing students with disabilities using adaptive methods has recently come to the forefront of educational dialogue due to the planned use of adaptive tests for all but a small percentage of students in the accountability tests being developed (Stone & Davey, 2011). The use of adaptive methods of testing for students with disabilities offers the opportunity to better measure the skills and abilities of these students, higher proportions of whom tend to obtain scores in the lower tail of the proficiency distribution. In addition, there may be improvements in motivation and engagement gained by administering a

test that is more appropriate in terms of difficulty. One concern that has been raised about adaptive testing of students with disabilities is that the resulting tests may not cover only grade-level content (Minnema et al., 2000). This is both a theoretical consequence and a practical one for the majority of students who are at grade-level proficiency, although the item pool can be chosen with the goal of creating a grade-level appropriate assessment. However, developing a content-constrained pool that robustly covers the required proficiency range may indeed be challenging in practice, and it may still not cover the range of abilities spanned by test takers. Note that this latter concern is potentially true for any test. MSTs again add some measure of control in this area by allowing partial preadministration assembly.

Because the focus of the field test was on assessment of students with learning disabilities, research questions about the characteristics of the routing test and routing decisions for this population of students were important secondary issues to investigate. This report begins with a description of the DARA field test participants and materials before proceeding to the analyses and results based on several research questions that we motivate here.

- Tests that include adaptive features (i.e., elements that are differentiated by individual or group) should provide an improvement over tests geared toward the norm of what may be a broad population, leading to increased engagement and motivation of test takers. Test-taker motivation was examined in this study through analysis of posttest survey responses.
- We were also interested in how the routing test functioned psychometrically. The routing test was a subset of items from the full criterion test, which was designed to be similar to state accountability tests, so we wanted to evaluate how performance on the routing test compared to that on the students' state test. Efficiency due to shorter tests is another selling point of adaptive testing. Because fatigue can be a factor when tests are too lengthy, we were interested in determining whether the routing test could have been shortened while preserving the quality of the decision (i.e., routing students consistently and reliably).
- We also investigated the stability of decisions based on the cut score by exploring what would have happened had a different cut score been chosen.

Therefore, the objective of this secondary analysis was to answer the following research questions:

1. What were students' perceptions of the field test?

2. What were the psychometric characteristics of the routing test and its passage subsets? How might the length and content of the routing test affect the quality of the routing decision?
3. How might the cut score associated with the routing test affect the quality of the routing decision?

We have organized our presentation around those main questions.

Descriptions of Tests and Participants

Participants

The two-stage field test was administered to 275 students with RLD and 486 students with no learning disability (NLD), in the eighth grade, from 26 schools in Massachusetts. The subset of students with RLD who had all relevant scores, as described in this section, was the focus of the present data analyses. Table 1 shows the percentages of students categorized into the proficiency categories on their state test both operationally (from the technical report of the state test) and for RLD students in the field test sample taking the accommodated (Component Test 1) and nonaccommodated (Component Test 2) Stage 2 tests. The state test information is included in order to provide a proficiency context for the sample of students participating in the study. The overall RLD sample appears similar in distribution to the students with disabilities taking the state assessment; however, the sample had a smaller percentage of students who would have been categorized as Proficient given their state test scores. It is important to note that the students with disabilities category used in the state's technical report may include other disability subtypes.

Table 1

Percentages of Students in State Test Proficiency Categories From the Official Administration and the Field Test (RLD)

		Warning/failing	Needs improvement	Proficient	Advanced
Operational	All students	7	18	63	12
	Students with disabilities	27	36	35	1
RLD field test participants	Overall	27	50	22	< 1
	Accommodated	42	46	11	0
	Nonaccommodated	12	55	33	1

Note. RLD = reading-based learning disabilities.

A total of 132 RLD students routed to Component Test 1 had scores on all relevant measures (criterion, Component Test 1, and ORF), and 130 RLD students routed to Component Test 2 had scores on all relevant measures (criterion and Component Test 2). Table 2 contains descriptive statistics for students in the study who took the accommodated test, and Table 3 contains analogous information for students taking the nonaccommodated test. The tables contain field test information only for the subset of RLD students who had completed all relevant field test pieces and for whom we had operational state test scores. We provide state test score summary information in Tables 2 and 3 to give an idea of how the RLD samples performed on a relevant external criterion.

Table 2

Descriptive Score Statistics for Students With Reading-Based Learning Disabilities With Full Data Who Were Routed to the Accommodated Test (Component Test 1, N = 132)

	Maximum possible score	M	SD
State test (scaled)	280	223.6	10.09
Routing (Part 1 of criterion)	32	9.4	2.82
Criterion (Parts 1 and 2)	48	14.9	4.07
Component Test 1 multiple choice	42	19.0	5.94
Component Test 1 scale	48	20.6	5.50

Table 3

Descriptive Score Statistics for Students With Reading-Based Learning Disabilities With Full Data Who Were Routed to the Nonaccommodated Test (Component Test 2, N = 130)

	Maximum possible score	M	SD
State test (scaled)	280	233.2	10.83
Routing (Part 1 of criterion)	32	18.7	3.82
Criterion (Parts 1 and 2)	48	27.3	6.02
Component Test 2 multiple choice	42	22.6	7.45
Component Test 2 scale	48	25.9	8.51

Test Materials

The field test routing measure was designed to route RLD students to an accommodated or nonaccommodated test based on their scores on that measure. The full field test consisted of the administration of two tests, both of which were assembled based on classical item statistics

that resulted from a pilot administration. Both tests were designed to measure grade-level constructs in English language arts with a focus on reading comprehension. The data collection structure had three steps. The design is presented in Figure 1 and is represented in a scaled-down version in Figure 2 for reference, followed by an overview of the test elements. A more thorough description of these elements follows in the next section.

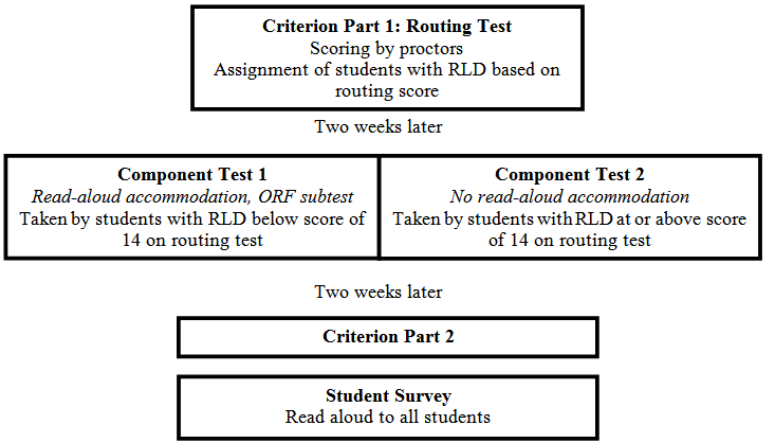

***Figure 2*. Data collection structure; RLD = reading learning disabilities.**

The criterion test, constructed to represent a typical state assessment, was administered in two parts. Part 1 of the criterion test also served as the routing test for the study. The component test, a prototype test designed to be accessible to students with RLD, had two slightly different versions that were referred to as Component Test 1 and Component Test 2. First, the routing test (Part 1 of the criterion test) was administered to all students and, then, proctors scored the routing test. Students with RLD achieving a score below the designated number-right passing score (14 items correct out of 32 items) were routed to an accommodated test (Component Test 1), and students with RLD who had scores surpassing the cut were routed to a nonaccommodated test (Component Test 2).

Unlike traditional MSTs, both second-stage tests consisted of the same MC items, and scores from the two stages were not combined; however, the accommodated test was administered with a read-aloud accommodation and included an ORF subtest to allow students to demonstrate proficiency in the aspects of reading that may be masked by the read-aloud accommodation. This explicit difference in conditions and implicit difference in difficulty between the two second-stage tests is one reason that we use the terms condition-adaptive and

multistage to describe the test. Additionally, the second-stage test to which a student was assigned depended upon performance on the first-stage routing test, much like the dependencies between stages in a typical multistage assessment. Part 2 of the criterion test was administered last to all test takers and consisted of 16 items.

Item and test specifications for the criterion test and the two second-stage tests, Component Tests 1 and 2, are shown in Tables 4 and 5. All tests were designed using the National Assessment of Educational Progress 2009 Reading Framework (National Assessment Governing Board, 2008). While the criterion test approximated a typical state assessment, the second-stage tests were both created to adhere to the *Accessibility Principles for Reading Assessments* (Thurlow et al., 2009). For a comprehensive description of the pilot test and field test administration and results, see Laitusis et al. (in press).

Table 4

Specifications for the Criterion Test

Passage type	Item type	No. of items
	Part 1	
Expository /informational (2 passages)	Critique/evaluate	3
	Integrate/interpret	8
	Locate/recall	5
Literary (2 passages)	Critique/evaluate	4
	Integrate/interpret	8
	Locate/recall	4
Total items in Part 1		32
	Part 2	
Expository /informational (1 passage)	Critique/evaluate	2
	Integrate/interpret	3
	Locate/recall	3
Literary (1 passage)	Critique/evaluate	2
	Integrate/interpret	4
	Locate/recall	2
Total items in Part 2	Critique/evaluate	16
Total items in criterion test		48

Table 5

Specifications for Component Tests 1 and 2

Passage type	Item type	Number of items
MC reading comprehension		
Expository /informational (3 passages)	Critique/evaluate	4
	Integrate/interpret	11
	Locate/recall	6
Literary (3 passages)	Critique/evaluate	6
	Integrate/interpret	9
	Locate/recall	6
Total MC items in each component test		42
Oral reading fluency subtest (excerpts from the four passages from the criterion test, Part 1)		
Expository /informational (2 passages)	N/A	
Literary (2 passages)	N/A	

Note. MC = multiple choice.

All passages were selected from released reading portions of accountability assessments from several states. Because the passages had previously been used in operational testing, they had items that were likely to have been piloted or pretested by the states prior to operational use. Additional items that were required were developed by ETS test developers. In a June 2009 pilot, which included students with learning disabilities, all items to be used were pretested. There were 11 three-passage forms (16 total passages, with some overlap to provide common items for linking) administered to NLD and RLD students, and the forms were equated in sequence with the first form as the base form using mean-sigma equating. Passages and items were selected based on student performance and relevant psychometric statistics (e.g., P+, point-biserial correlation, and the proportion of students who did not respond to the item) as well as an item-by-item review by researchers. The review was undertaken to identify directions, passages, or items that had accessibility issues (e.g., the options contained too much text, or the items required referring back and forth to the passage) or inconsistencies in format. We converted P+ values to the delta scale,[1] a common metric used at ETS for test development. This conversion resulted in lower delta values representing easier items, whereas items with lower P+ values are more difficult. The delta scale has a mean of 13 and a standard deviation of 4. Items were considered for use in various parts of the field test based on those criteria. Table 6 contains the

summary statistics for the equated-delta difficulty of the items selected and assembled for the criterion and component tests, based on the June 2009 pilot.

Table 6

Field Test Difficulty as Assembled

	Equated delta	Criterion	Component
Section 1	No. of items	32	21
	Mean	12.9	12.1
	SD	0.7	0.7
	Min	12	11.2
	Max	14.8	13.8
Section 2	No. of items	16	21
	Mean	12.6	12.2
	SD	1	1.5
	Min	10.7	9
	Max	14	15.2

Routing Mechanism

The desired outcome of the routing was to achieve separation between high- and low-performing RLD students, so that the most appropriate Stage 2 test could be assigned, while maintaining adequate sample sizes in both groups. To achieve these goals, the decision was made to route approximately 40% of the RLD students to the accommodated test. An initial passing score was set by examining 30th, 40th, and 50th percentiles of the RLD student scores on similar material (i.e., a form containing three of the four passages from the routing test) from the pilot. The passing score was modified after evaluation of preliminary data from four schools that were diverse in terms of average state test performance. Based on this evaluation, RLD students with scores of 14 and above were designated to be routed to Component Test 2, and RLD students with scores of 13 and below were routed to Component Test 1.

Administration of the Tests

The DARA field test data collection was carried out in February and March of 2010. For ease of administration, all test components were printed in the same test book; however, students only accessed the current test section and were not allowed to visit other test sections at that time. Students answered all questions by circling the correct answers in their test books. Students were given the same test book at three separate time periods: first, when Part 1 of the criterion test was administered; second, when either Component Test 1 or 2 was administered; and, third, when Part 2 of the criterion test was administered and the student survey was to be completed.

Administration of criterion test Part 1 (routing test). The routing test was administered in intact classrooms and was then scored by the proctors. Based on their scores, RLD students were assigned to Component Test 1 or Component Test 2.

Administration of second-stage tests (Component Tests 1 and 2). Approximately 2 weeks after the first stage was completed, Component Tests 1 and 2 were administered.

- Component Test 1 was administered with a read-aloud accommodation available via MP3. The read-aloud accommodation was provided as separate tracks on MP3 players, allowing students to independently navigate through the parts of the passages that they needed to have read aloud or repeated. An ORF subtest was also administered as part of that test in order to directly assess the component skill of decoding.
    - **Administration of the ORF measure.** The ORF test in Component Test 1 was composed of excerpts (each was 300–400 words in length) from the four routing test passages and was delivered on the computer. This portion of the test was administered to up to 15 students at a time. Students read aloud each passage as quickly and accurately as possible using a headset with a microphone.
- Component Test 2 did not allow for audio presentation of the content and did not include the ORF subtest.

Administration of Criterion Test Part 2. Criterion Test Part 2 was administered approximately 2 weeks after the administration of Component Tests 1 and 2.

Administration of student survey. The student survey, capturing students' perceptions of the assessment, appeared at the end of the test booklet and was read aloud to all students. The survey questions are included in the appendix of this report.

Research Question Methodology and Results

Student Perceptions of the Field Test

Method. One of the aspects of adaptive testing that is often claimed to be a benefit for the test taker is increased motivation and engagement. This and other features of the testing experience were evaluated by analyzing the student survey response frequencies by routing test score for RLD students.

Results

We include summary tables collapsed over routing scores. Full tables including results by individual routing scores are available in Laitusis et al. (in press). Tables 7–11 contain response frequencies for questions asked of all participants. Only RLD responses are tallied because only that group was routed rather than randomly assigned. Scores are collapsed across categories. Because one focus was on students on the cusp of the cut score, we group students as those more than 2 points below the cut score of 14 (i.e., routing scores of 0–11), students within 2 points of—but below—the cut score (i.e., routing scores of 12 and 13), students within 2 points of—but above—the cut score (i.e., routing scores of 14 and 15), and students more than 2 points above the cut score (i.e., routing scores of 16–32). Table 12 also contains only RLD responses, and the participants are grouped into those routed to Component Test 1 (routing score of 13 or below) and those routed to Component Test 2 (routing score of 14 or above), rather than by individual routing score. Tables 13–15 only contain responses from RLD participants who were routed to Component Test 1 and received the read-aloud accommodation.

Tables 7 and 8 display the frequencies of student responses, by routing score, to questions of how their reading comprehension proficiency compares to that of other eighth graders when they have material read to them and when they read it on their own (respectively). It is interesting to note that, in both conditions, most students said that their level of understanding is the same as that of other eighth graders. Relatively few students (8% and 19%, respectively) said that their understanding was worse than that of their peers, and the score distribution for the students choosing that option was fairly widespread rather than concentrated at the lower scores. These results point to the issues of self-awareness and awareness of proficiency in relation to one's peers.

Table 7

Survey Question 1: Compared to Other Eighth Graders in Your School, How Well Do You Think You Understand What Is Read Aloud to You?

Routing score category	Better than other eighth graders	The same as other eighth graders	Worse than other eighth graders	[Omitted]	Total
More than 2 points below cut	20	60	14	5	99
Within 2 points below cut	7	31	2	2	42
Within 2 points above cut	10	17	1	1	29
More than 2 points above cut	21	78	4	2	105
Total	58	186	21	10	275

Table 8

Survey Question 2: Compared to Other Eighth Graders in Your School, How Well Do You Think You Understand What You Read on Your Own?

Routing score category	Better than other eighth graders	The same as other eighth graders	Worse than other eighth graders	[Omitted]	Total
More than 2 points below cut	21	51	22	5	99
Within 2 points below cut	8	26	6	2	42
Within 2 points above cut	4	20	4	1	29
More than 2 points above cut	24	59	20	2	105
Total	57	156	52	10	275

When asked how hard they tried on this test as compared to other tests (Table 9), most students across the continuum (69%) said that they tried about the same. Students at the lower end of the routing test scale were more likely to say that they tried harder on this test than that they tried about the same or not as hard. Most students (73%) said that they would prefer to take a reading test with a MP3 player (Table 10). It may appear surprising that 23% of students claimed to prefer to take the test without the additional MP3 assistance; however, it should be noted that students choosing that option were overwhelmingly above the cut score, meaning that they did not experience the Component Test 1 test with the read-aloud accommodation as a comparison and may not have access to read-aloud assistance during routine instruction or assessment. In addition, their reading speed is likely to be faster than the rate at which the text was read aloud.

Table 9

Survey Question 3: Did You Try as Hard on These Tests as You Do on Other Tests That You Take?

Routing score category	No, I didn't try as hard as on other tests	Yes, I tried about as hard as on other tests	Yes, I tried harder than on other tests	[Omitted]	Total
More than 2 points below cut	14	61	19	5	99
Within 2 points below cut	1	33	6	2	42
Within 2 points above cut	4	17	7	1	29
More than 2 points above cut	10	78	15	2	105
Total	29	189	47	10	275

Table 10

Survey Question 4: If Given the Choice of Taking a Reading Test With or Without an MP3 Player to Read the Test to You, Which Would You Prefer?

Routing score category	With an MP3 player to read aloud the test in addition to the paper copy	Without an MP3 player to read aloud the test	[Omitted]	Total
More than 2 points below cut	86	8	5	99
Within 2 points below cut	31	9	2	42
Within 2 points above cut	14	14	1	29
More than 2 points above cut	70	33	2	105
Total	201	64	10	275

Table 11 shows how many RLD students, by routing score, thought that they did better on the first day (routing part of criterion) test or second day (component) test. Of students who were routed to the accommodated test (Component Test 1), approximately 67% thought that they did better on the second day (component) test. However, for students routed to the nonaccommodated test (Component Test 2), approximately 66% felt that they did better on the first day (routing part of the criterion). This is in line with the finding (see Laitusis et al., in press) that RLD students experienced a boost in score (component minus criterion) when taking the accommodated version, but that there was a small decrease in mean score for students routed to Component Test 2. This may indicate that the accessibility efforts in the Component Test 2 design were not necessarily helpful for some readers.

Table 11

Survey Question 5: Which Test Do You Think You Did Better on?

Routing score category	The first test I took (1st day of testing)	The second test I took (2nd day of testing)	[Omitted]	Total
More than 2 points below cut	28	65	6	99
Within 2 points below cut	11	29	2	42
Within 2 points above cut	19	9	1	29
More than 2 points above cut	69	34	2	105
Total	127	137	11	275

Table 12 shows the categories of suggestions by students about how to make a better reading test for them, and it is divided by the component test that was taken. Most students had

no suggestions. Of those who made suggestions, the read-aloud accommodation was the overwhelming choice. Students routed to Component Test 2, who did not use the read aloud during the test, seemed more likely to suggest changes to the content or difficulty of the test than were students routed to Component Test 1. Most of the students in that group who did suggest read aloud had scores of 19 or less on the routing test (73%) and were, therefore, on the lower end of the Component Test 2 group. This may indicate that those students typically receive read aloud in instruction or when taking assessments and that the Component Test 1 test may have been more appropriate for them. One student suggested that students be able to individually adjust the speed on the MP3. Some students appeared to think that the survey question was asking what they should do in order to improve their performance on the test (see Strategies category).

Table 12

Survey Question 6: Do You Have Any Ideas About How to Make a Better Reading Test for You? (Reading-Based Learning Disabilities)

Routing result	No suggestions[a]	Read aloud[b]	Strategies[c]	Timing[d]	Content[e]	Difficulty[f]	Help[g]	Administration[h]
Component Test 1	55	40	10	2	4	8	1	3
Component Test 2	68	26	8	7	13	12	2	2
Total	123	66	18	9	17	20	3	5

[a]No suggestions: no suggestions were noted. [b]Read aloud: anything having to do with read aloud, whether the student suggested that read aloud be available on electronic medium or individually or group administered by a teacher or proctor. [c]Strategies: any test-taking strategies that students thought would make the test better, such as being allowed to chew gum, listening to music, studying hard, highlighting pertinent information, or reading the questions before the passage. [d]Timing: suggestions related to shorter testing sessions, more breaks, time for stretching, more time allowed per section, and no time limits. [e]Content: suggestions about including shorter or more interesting (or contemporary) stories or adding pictures. [f]Difficulty: refers to the format in which questions were asked (e.g., use only MC questions, ask in the form of a crossword puzzle, list the questions before the passage), the difficulty of the questions asked, the length of the test, and the format of the test (e.g., less text per page). [g]Help: suggestions that a dictionary, glossary, or additional help be provided. [h]Administration: contains suggestions such as administering all of the test on computer, administering the test in a quieter room without interruptions, or paying students for taking the test.

Table 13 shows how much of the test was listened to on MP3 by Component Test 1 test takers. A majority, 70%, said that they had listened to all of the test on MP3. One person with a routing score of 10 claimed not to have listened to any of the test on MP3. In Table 14, the results from the same group being asked which test they thought they did better on (without or with read aloud) are displayed. About 76% of the group said that they thought that they had performed better on the test with read aloud. Table 15 shows results from the same group being asked whether or not they had had a test read to them before. The majority (57%) had experienced having teachers read tests to them, about 12% had listened to tests on MP3, and about 18% had not had a test read to them before.

Table 13

Survey Question 7: How Much of the Test Did You Listen to?

Routing score category	All	Most	Some	None	[Omitted]	Total
More than 2 points below cut	67	18	11	1	2	99
Within 2 points below cut	31	4	3	0	4	42
Total	98	22	14	1	6	141

Table 14

Survey Question 8: Which Test Do You Think You Did Better on?

Routing score category	The test that was read aloud by the MP3 player	The tests that I read to myself	About the same on all the tests	[Omitted]	Total
More than 2 points below cut	79	8	10	2	99
Within 2 points below cut	28	6	4	4	42
Total	107	14	14	6	141

Table 15

Survey Question 9: Have You Ever Had a Test Read Aloud to You?

Routing score category	No	Yes, read by MP3 player	Yes, read by CD player or tape player	Yes, read by a teacher	Yes, read by computer (e.g., Kurzweil or TextHelp)	[Omitted]	Total
More than 2 points below cut	19	10	6	59	2	3	99
Within 2 points below cut	6	7	3	22	0	4	42
Total	25	17	9	81	2	7	141

Routing Test Content, Psychometric Measures, and Outcomes

Method

Cronbach's alpha was calculated to determine the internal consistency reliability of the routing test. The statistic is computed as

$$\frac{p}{p-1}\left(1-\frac{\sum_{j=1}^{p} Var(X_j)}{Var(X)}\right),$$

where p is the total number of items, Xj represents the score on the jth item, and X represents the total score on all p items.

Reliability (alpha) was also calculated for all subsets of passages to provide evidence of whether a shorter routing test would have been able to provide the same quality routing.

Correlations with external criteria (e.g., the score from the state assessment) were computed to provide an indication of how well the routing test approximated a typical state assessment. The state assessment has a score scale ranging from 200–280; scores for some students were provided by the schools. Correlations were also computed between the routing test score and the component test scores.

It is important to note that the routing procedure led to a direct restriction of range on the routing test score for students taking the different component tests. This restricted sample selection typically leads to an underestimation of the correlation between a predictor and criterion in the overall population, and a correction mechanism is sometimes used to offset the bias. One proposed correction for the correlation was that offered by Lord and Novick (1968, p. 143) in population terms and repeated in the sample terms used here in Gross and Kagen (1983). Let X be the routing scores in the full RLD group, XR be the routing scores in the RLD group routed to Component Test 1, and Y be the Component Test 1 scores. Then,

$$r_{XR,Y} = \frac{s_{XR,Y}}{s_{XR} \cdot s_Y}$$

represents the uncorrected correlation between the restricted criterion score and the component score, where $s_{XR,Y}$ is the covariance of *XR* and *Y*, s_{XR} is the standard deviation of *XR*, and s_Y is the standard deviation of *Y*. Using the correction formula,

$$r_{X,Y} = \frac{r_{XR,Y}}{\sqrt{r_{XR,Y}^2 + \frac{s_{XR}^2}{s_X^2}(1 - r_{XR,Y}^2)}}$$

represents the corrected correlation. An analogous procedure was used to compute the corrected correlation involving Component Test 2.

Content and Routing Decision Analysis

First, we evaluated the relationship of the actual routing test score to the final component test scores for both groups of RLD students. For students taking Component Test 1, the final Component Test 1 score was a weighted combination of the reading comprehension MC test score and the ORF subtest that had been scaled, for the purpose of comparison of the final Component Test 1 score and the criterion test, to a total of 48 points (the maximum number of points on the criterion test). Laitusis et al. (in press) included a more detailed description of the computation of weights and the scaling procedure.

The impact of the routing test content and length was investigated by evaluating the psychometric characteristics and routing percentages for all passage subsets of the criterion test (Parts 1 and 2). Recall that the cut score used in the field trial for the four routing test passages was 14; in other words, students achieving a score of 13 or below were routed to the accommodated second-stage test. The cut score for each potential routing subset was set by multiplying its test length by a factor of 0.40625. This factor was arrived at by dividing 13 by the number of items in the original routing test (32) to determine what proportion of total score on the routing test was the maximum score for which students would be routed to Component Test 1. Thus, passage subsets with varying lengths were evaluated using proportional cut scores (e.g., for one passage with eight items, 0.40625 * 8 = 3.25 items, which was rounded to a score of 3). Although only the first four passages were chronologically appropriate for routing during the field test, we also evaluated the two Criterion Part 2 passages to get a sense of whether the routing test was composed of the optimal set of passages. The full criterion test was administered

to all students; therefore, the portion that was chosen for routing is just one subset of passages that could have been selected from the full set of six. Although Criterion Part 1 and Criterion Part 2 were administered on different occasions, the administrations were not separated by enough time that a significant growth in achievement would be expected.

For a total of six passages, there were 63 possible subsets of passages that could have been used to route the RLD students had the passages all been delivered prior to the second-stage assignment. Because the subsets had different lengths (one, two, three, four, five, or six passages with 8, 16, 24, 32, 40, or 48 items, respectively), we used the Spearman-Brown prophecy formula to adjust the reliabilities to those expected for 48 items for all subsets. Despite some criticisms in the literature of the formula's use (e.g., Charter, 2001), the vastly different subset lengths made an adjustment appropriate. The Spearman-Brown formula can be written as

$$\alpha_{S-B} = \frac{\frac{k_2}{k_1} \cdot \alpha}{\left(1 + \left(\frac{k_2}{k_1} - 1\right) \cdot \alpha\right)},$$

where α represents the regular Cronbach alpha, k_2 is the number of items in the maximum test length (in this case), and k_1 is the number of items in the actual test under consideration.

In addition to the psychometric analyses just described, we compared the percentages of students routed to each second-stage test using each passage subset to the percentages obtained by the actual routing test.

Results

Table 16 contains the correlations between the state test, routing test, and component test MC portion for those RLD students routed to the accommodated test (Component Test 1). For this group, *component scale* refers to the scaled and weighted combination of the MC and ORF sections. The state test information is included for two reasons. First, the state test serves as an external validity criterion. Second, the criterion test was designed to approximate a typical state test; therefore, it was of interest to relate performance on the state and criterion tests if possible. Table 17 displays analogous results for the RLD students routed to the nonaccommodated test (Component Test 2). In this group, component scale is a linear transformation of the MC section score to a possible 48 points.

Table 16

Test Score Correlations for Reading-Based Learning Disabilities (RLD) Group Routed to Accommodated Test

	State test	Routing	Component MC	Component Test 1 scale
State test	1	0.22	0.42	0.46
Routing (Criterion Part 1)		1	0.17	0.23
Component Test 1 MC			1	0.94
Component Test 1 scale				1

Note. MC = multiple choice.

Table 17

Test Score Correlations for Reading-Based Learning Disabilities (RLD) Group Routed to Nonaccommodated Test

	State test	Routing	Component Test 2 scale
State test	1	0.49	0.64
Routing (Criterion Part 1)		1	0.66
Component Test 2 scale			1

If the correction formula, previously given, is applied to the correlations between the routing and component tests, the corrected correlations are (predictably) larger. For the accommodated test, the correlation between routing and MC section is 0.33 (versus 0.17), and between routing and scale score is 0.44 (versus 0.23). For the nonaccommodated test, the corrected correlation between the routing test and the Component Test 2 scale is 0.78 (versus 0.66). It is interesting to note that the correlation of the routing test and the state test, while positive, is low in this group. However, the state test has 84 items (including some open-ended items), while the routing test has only 32 items. The routing test reliabilities (uncorrected for restricted range) were 0.20 in the group routed to the accommodated test and 0.54 in the group routed to the nonaccommodated test. It should be noted that the routing test reliability for the combined group of all RLD students taking the field test was 0.79 (Laitusis et al., in press).

It is evident from the scatterplot in Figure 3 that there is a positive relationship between the routing and component tests overall, and it appears as if there is a stronger positive relationship for the students taking Component Test 2.

The results of the routing test length and content evaluation are included in Table 18.

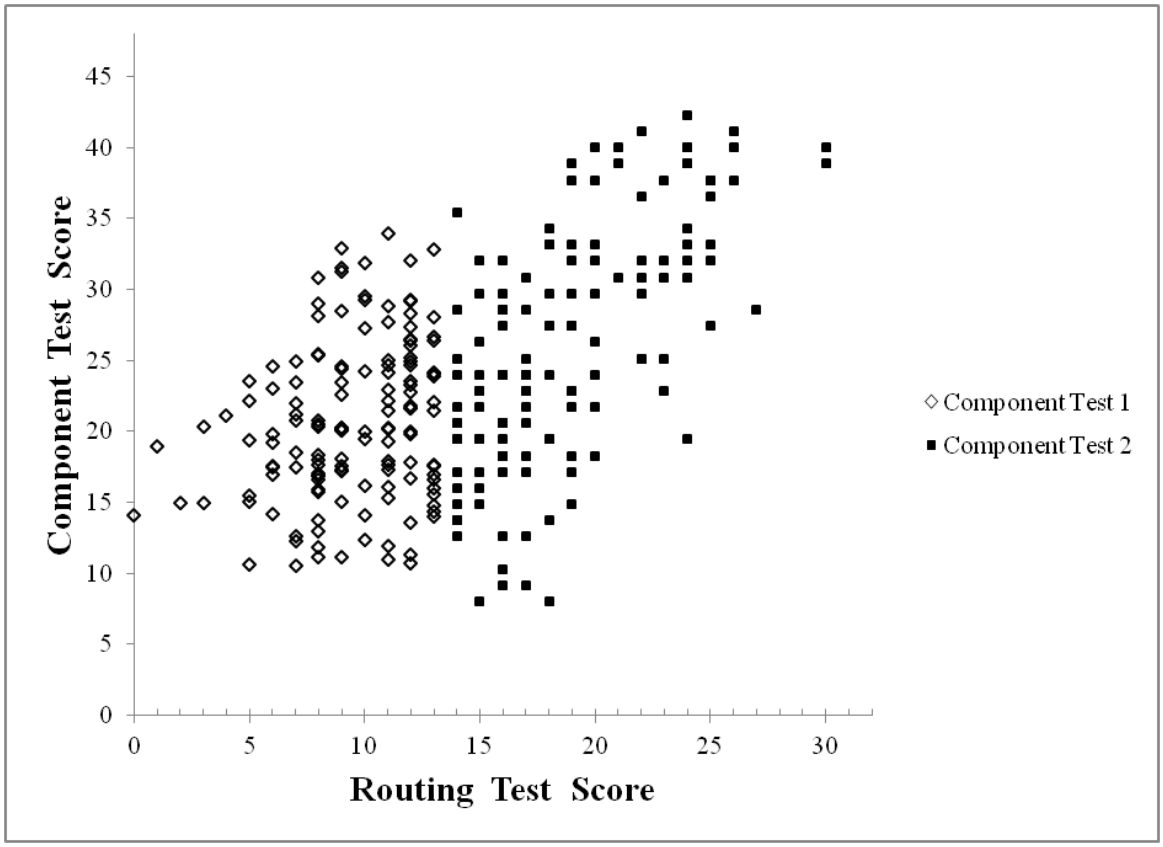

***Figure 3*. Component test versus routing test scores for students with reading-based learning disabilities (RLD).**

Table 18

Median Alphas and Routing Decision Percentages for Criterion Passage Subsets of Varying Lengths

				Routed to C1 by this subset			Routed to C2 by this subset		
Number of passages	Number of subsets of this length	Alpha	Spearman-Brown alpha	% out of all 262 RLD	% agreeing with original decision out of 262 RLD	% agreeing with original decision out of 132 originally routed	% out of all 262 RLD	% agreeing with original decision out of 262 RLD	% agreeing with original decision out of 130 originally routed
1	6	0.552	0.880	52	37	73	48	34	69
2	15	0.672	0.860	60	45	89	40	35	70
3	20	0.736	0.848	55	44	88	45	40	80
4	15	0.783	0.844	51	45	89	49	41	83
5	6	0.819	0.845	48	44	87	52	46	93
6	1	0.841	0.841	51	46	92	49	45	90

Note. RLD = reading-based learning disabilities.

These results indicate that there are a number of ways in which routing adequacy can be evaluated.

- Absolute reliability: Perhaps unsurprisingly, the winner in this category is the full criterion test (6 passages, alpha = 0.841).
- Adjusted reliability: Perhaps surprisingly, Passage 2 has the highest adjusted reliability using the Spearman-Brown formula (0.916), followed by Passage 1 (0.901).
- Percentage routed to Component Test 1: Our goal was to route approximately 40% of students to Component Test 1. Passage 2 routed 41% of students there, and Passage 1 routed 42% of students there.
- Percentage in agreement with previous routing decision: If we assume that the original routing test with Passages 1–4 performed its job well, then we might want to look at how many test takers the subsets correctly routed to the same component.
    - What percentage of people originally routed to Component Test 1 again were routed to Component Test 1: The closest match to the routing to Component Test 1 of the original routing test was the subset composed of Passages 2–4 (i.e., the full routing test minus Passage 1). This subset correctly routed 98% of the original 132 Component Test 1 test takers to Component Test 1 again, and this was a total percentage of 49% correctly rerouted to Component Test 1 in the full sample (i.e., all 262 RLD students).
    - What percentage of people originally routed to Component Test 2 again were routed to Component Test 2: The closest matches to the routing to Component Test 2 of the original routing test were the subsets composed of Passages 1, 2, 4, and 5, and 1, 2, 4, and 6. These subsets each correctly routed 100% of the original Component Test 2 test takers to Component Test 2 again, and this was a total percentage of 50% correctly rerouted to Component Test 2 in the full sample. However, it should be noted that each of these subsets seemed to overroute test takers to Component Test 2 compared to the original routing, as they routed 82% of all test takers to that component test (recall that the target was approximately 60%).
- Correlations with component test scores: Table 19 contains the median uncorrected and corrected (for restriction of range) correlations of the different possible routing tests with

the component tests. An examination of the information in this table shows that the relationship between criterion passages and component scores was greater for the RLD students who took Component Test 2 (i.e., the students who had been routed to that nonaccommodated test due to higher routing test performance). The individual passages varied considerably in their correlations with the various scores given. If one focuses on the corrected correlations between the individual passage scores and the total test scores, the individual-passage correlations with total score ranged from 0.02–0.38 for Component Test 1 and from 0.40–0.57 for Component Test 2. For Component Test 1, the maximum correlation with total test score was 0.66 and was obtained by a five-passage subset. For Component Test 2, the maximum correlation between subset and total test score was 0.85, and this was obtained by the six-passage subset and by two of the four-passage subsets.

Table 19

Median Correlations Between Fixed-Length Criterion Test Passage Subsets and Component Test Scores

Number of passages	Number of subsets of this length	Component Test 1			Component Test 2	Component Test 1			Component Test 2
		Multiple choice	Fluency	Total	Total	Multiple choice	Fluency	Total	Total
1	6	0.22	0.13	0.23	0.45	0.29	0.16	0.30	0.52
2	15	0.17	0.17	0.24	0.58	0.26	0.23	0.36	0.65
3	20	0.25	0.20	0.30	0.66	0.40	0.32	0.48	0.73
4	15	0.32	0.25	0.37	0.71	0.51	0.43	0.57	0.79
5	6	0.29	0.25	0.36	0.74	0.50	0.46	0.59	0.82
6	1	0.33	0.27	0.40	0.76	0.57	0.48	0.65	0.85

Cut Score Sensitivity

Method. The cut score used operationally was chosen on the basis of testing goals and pilot test and preliminary field test data, but the score was inevitably somewhat arbitrary. It was of interest to determine, using the set of 32 items that were actually used for routing, to what extent moving the cut score would change the percentage of students routed. To examine the stability of the cut score, routing agreement was evaluated between the chosen cut score and a range of other possibilities for the score used with a total test length of 32 items.

Results. Figure 4 shows the percentage of RLD students who would be routed to Component Test 2 using the original routing test (Passages 1–4) with various passing scores. The plot suggests that relatively minor perturbations of the score distribution would have altered the percentages of students routed to the two component tests.

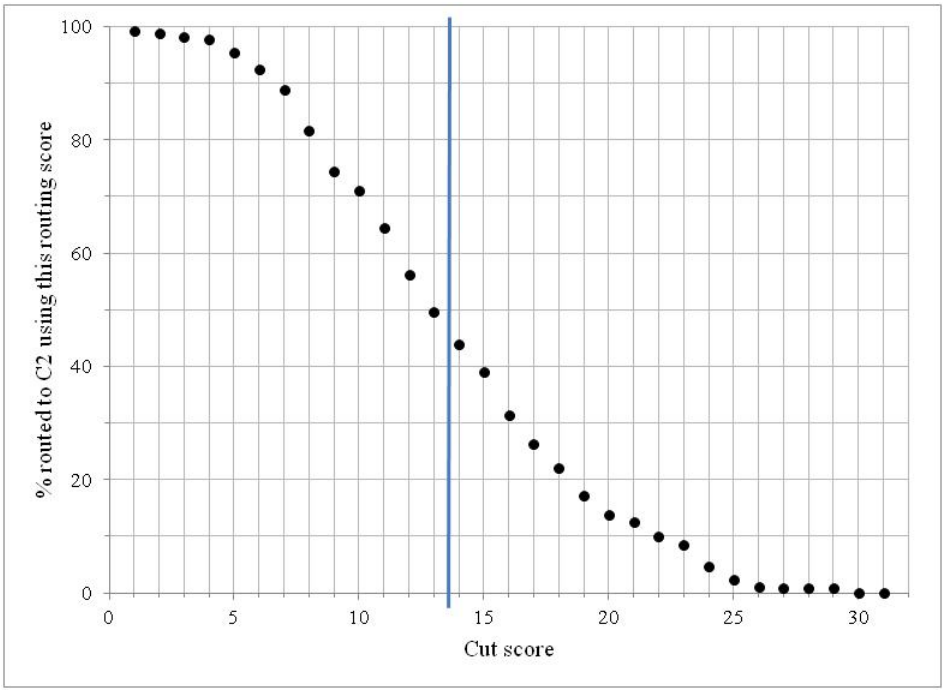

Figure 4. **Impact of various cut scores on percentage of students with reading-based learning disabilities (RLD) routed to Component Test 2.**

Discussion

In this paper, we explored the use of an adaptive testing alternative to the typical linear paper and pencil tests that are currently used for most state accountability assessments. The exploration of this alternative is particularly timely, because many state accountability programs are currently considering moving from linear paper and pencil assessments to some type of computerized adaptive assessment. Finding alternatives to linear paper and pencil assessments is important for a number of reasons including the use of these alternatives for certain populations such as individuals with RLD.

The adaptive test we explored in this study was not a typical MST in that it did not route test takers to different groups of items or testlets (i.e., stages that resulted in a score that was formed by combining scores across the stages). However, the structure of the test we investigated

does provide an adaptive alternative to the conventional fixed-form linear tests that are used for accountability by most states. Whereas many states are currently faced with the choice of either (a) allowing a read-aloud modification for students who need it, possibly changing the construct, and not being able to count the score for accountability or (b) not allowing the modification and essentially losing the chance to measure comprehension for those students, the two-stage, condition-adaptive format we studied would allow for componentwise measurement of reading (decoding, comprehension) that might improve measurement quality, increase student engagement, and increase inclusion in accountability measures. In addition, the component approach described in this study provides students, teachers, and parents with additional information that might be used for diagnostic purposes; for example, information would be available about a student's decoding skills and how these skills relate to the student's ability to comprehend printed text with a read-aloud accommodation.

The study sought to examine the routing test used as part of the adaptive process in detail. We explored the content and psychometric characteristics of the routing test and also compared it to an external criterion (the state test score) and second-stage test scores. We evaluated the routing test performance by comparing the percentage of test takers that had been routed to either Component Test 1 or 2 (using the original routing test) to the percentage routed using the routing tests created from other available passages for the criterion test. We found that, in examining all criterion passage subsets for routing adequacy, it appeared that the routing test used in the field test performed similarly to the best of the criterion passage subsets. One goal of using the routing test in this study was to ensure adequate sample sizes by choosing target routing percentages for assignment to the two second-stage tests. Preliminary analysis of data from a representative group of schools indicated that our initial cut score was too low, in that too few students would be projected to be routed to the accommodated test. The cut-score sensitivity analysis on the full set of data indicated that moving the cut score just a few points in either direction could have, in this sample, had a larger effect than would be optimal on the percentages of students routed to the two component tests. This emphasizes the importance of (a) taking a very careful approach to setting cut scores a priori and (b) evaluating preliminary data, if possible, to get an idea of how well the routing procedure is working in the context of the metrics set forth by the researchers or test developers so that it can possibly be adjusted. Overall, the routing test used in the field trial seems to have been of reasonable quality and structure.

Another important finding of the study was the low correlations between the routing and state test scores. This finding requires further investigation. One possibility is that the state test scores may have errors in the reporting (as they were gathered by the test administrators and were not sent officially by the state); it is also possible that the scores may have been from tests given under different testing conditions (e.g., with a read-aloud modification or other test changes), the testing blueprints may not have been an adequate match, or may have been due to an all too common case of less motivation on a non high-stakes test. Because one of the goals of the study was to investigate an assessment that realistically reflected the characteristics of a typical state assessment, these hypotheses should be pursued.

An important finding of the study was that it appeared as though the multistage format of the assessment appealed to test takers with RLD. Although 11% of RLD participants claimed not to have tried as hard as on other tests, the vast majority stated that they had given the same amount of or more effort. In addition, students scoring below the cut score on the routing test reported that they preferred to take the test with the read-aloud accommodation provided by the MP3 player.

The ongoing educational policy reform has opened the door to possibilities for improving the assessment of all students including students with RLD. With the interest in adaptive testing for accountability purposes, it is important to examine how such tests might work for these students and for other students with disabilities. The adaptive nature can provide benefits for students who are performing in the tails of the proficiency distribution, but further investigation is required before some of the issues associated with using adaptive models with students with disabilities are resolved. In the case of the model used for this study, a key issue is how to provide a routing test that is on grade level but that is matched well enough to the ability level of the target group to provide scores that are reliable enough for routing purposes. We believe that given a large enough pool of items to use for test construction that this obstacle can be overcome and that the benefits of this type of assessment will well outweigh any additional costs.

References

Betz, N. E., & Weiss, D. J. (1976). *Psychological results of immediate knowledge of results and adaptive ability testing* (Research Report No. 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Charter, R. A. (2001). It is time to bury the Spearman-Brown "prophecy" formula for some common applications. *Educational and Psychological Measurement, 61,* 690–696.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing* (Measurement and Research Department Reports 2001-1). Arnhem, The Netherlands: Citogroep.

Goldstein, L. F. (2003). Spec. ed. tech sparks ideas. *Education Week, 22*(35), 27–29.

Gross, A. L., & Kagen, E. (1983). Not correcting for restriction of range can be advantageous. *Educational and Psychological Measurement, 43*(2), 389–396.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44–52.

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2002, April). *Comparison of the psychometric properties of several computer-based test designs for credentialing exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/NCEO/Onlinepubs/StateGuideUD/UDmanual.pdf

Johnstone, C. J., & Thurlow, M. L. (2010). Statewide testing of reading and possible implications for students with disabilities. *The Journal of Special Education, 46*(1), 17–25.

Kingsbury, G. G., & Hauser, C. (2004, April). *Computer adaptive testing and No Child Left Behind*. Paper presented at the annual meeting of the American Educational Research Association, San Diego CA. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ki04-01.pdf

Laitusis, C. C., Stone, E., Steinberg, J., & Cook, L. L. (in press). *Designing accessible reading assessments field test analyses final report*. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement 8*(3), 147–151.

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, *36*, 227–242.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test* (Research Bulletin No. RB-74-30). Princeton NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Minnema, J., Thurlow, M., Bielinksi, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis* (Out-of-Level Testing Project Report 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/OOLT1.html

National Assessment Governing Board. (2008, Spring). *Reading framework for the 2009 National Assessment of Educational Progress.* Washington, DC: Author. Retrieved from http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading09.pdf

No Child Left Behind Act of 2001, 20 U.S.C. 6301 *et seq.* (2001).

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing.* (Unpublished doctoral dissertation). University of Massachusetts at Amherst.

REL West/WestEd. (2008). *Considerations in statewide implementation of computer-adaptive testing*. Retrieved from http://relwest-archive.wested.org/system/memo_questions/11/attachments/original/Computer_20adaptive_20testing_20June_202008_1_.pdf

Stone, E., & Bruce, K. (2010, October). *Oral reading fluency as part of an accessible reading assessment for students with learning disabilities*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Stone, E., & Davey, T. (2011). *Computer-adaptive testing for students with disabilities: A review of the literature* (Research Report No. RR-11-32). Princeton, NJ: Educational Testing Service.

Sunderman, G. L. (2006). *The unraveling of No Child Left Behind: How negotiated change transforms the law*. Cambridge, MA: The Civil Rights Project, Harvard University.

Thompson, T., & Way, W. D. (2007, June). *Investigating CAT designs to achieve comparability with a paper test.* Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements*. Thousand Oaks, CA: Corwin Press.

Thurlow, M., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments.* Minneapolis, MN: National Accessible Reading Assessment Projects.

Thurlow, M. L., & Larson, J. (2011). *Accommodations for state reading assessments: Policies across the nation*. Minneapolis: University of Minnesota, Partnership for Accessible Reading Assessment.

Trotter, A. (2003). A question of direction. *Education Week, 22*(35), 17–21.

Weissman, A., Belov, D., & Armstrong, R. (2007). *Information-based versus number-correct routing in multistage classification tests* (LSAC Research Report No. 07-05). Newtown, PA: Law School Admissions Council.

Zenisky, A. L. (2004). *Evaluating the effects of several multistage test design variables on selected psychometric outcomes for certification and licensure agreement.* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

Notes

[1] $\Delta = 13\text{-}4\varphi\text{-}1(p)$, with $\varphi\text{-}1$ the inverse normal distribution.

Appendix

Field Test Student Survey

1. Compared to other 8th graders in your school, how well do you think you understand what is read aloud to you?
   (A) Better than other 8th graders
   (B) The same as other 8th graders
   (C) Worse than other 8th graders
2. Compared to other 8th graders in your school, how well do you think you understand what you read on your own?
   (A) Better than other 8th graders
   (B) The same as other 8th graders
   (C) Worse than other 8th graders
3. Did you try as hard on these tests as you do on other reading tests you take?
   (A) No, I didn't try as hard as on other tests
   (B) Yes, I tried about as hard as on other tests
   (C) Yes, I tried harder than on other tests
4. If given the choice of taking a reading test with or without an MP3 player to read the test to you, which would you prefer?
   (A) With an MP3 player to read aloud the test in addition to the paper copy
   (B) Without an MP3 player to read aloud the test
5. Which test do you think you did better on?
   (A) The first test I took (1st day of testing)
   (B) The second test I took (2nd day of testing)

6. Do you have any ideas about how to make a better reading test for you?

Only for students who used the MP3 players

7. How much of the test did you listen to?
   (A) All of the test
   (B) Most of the test
   (C) Some of the test
   (D) I did not listen to the test.
8. Which test do you think you did better on?
   (A) The test that was read aloud by the MP3 player
   (B) The tests that I read to myself
   (C) About the same on all the tests
9. Have you ever had a test read aloud to you?
   (A) No
   (B) Yes, read by MP3 player
   (C) Yes, read by CD player or tape player
   (D) Yes, read by a teacher
   (E) Yes, read by computer (e.g., Kurzweil or TextHelp)