CRESST REPORT 785

Jamal Abedi
Seth Leon
Jenny Kao
Robert Bayley
Nancy Ewers
Joan Herman
Kimberly Mundhenk

ACCESSIBLE READING ASSESSMENTS FOR STUDENTS WITH DISABILITIES: THE ROLE OF COGNITIVE, GRAMMATICAL, LEXICAL, AND TEXTUAL/VISUAL FEATURES

JANUARY, 2011

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Accessible Reading Assessments for Students with Disabilities: The Role of Cognitive, Grammatical, Lexical, and Textual/Visual Features

CRESST Report 785

Jamal Abedi
University of California, Davis/CRESST

Seth Leon and Jenny Kao
University of California, Los Angeles / CRESST

Robert Bayley
University of California, Davis

Nancy Ewers
University of California, Davis

Joan Herman
University of California, Los Angeles/CRESST

Kimberly Mundhenk
University of California, Davis

January, 2011

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under PR/award number Q3036031101, as administered by the U.S. Department of Education, Office of Special Education Programs to NCEO/University of Minnesota, therein subcontracted to the University of California, Davis.

To cite from this report, please use the following as your APA reference: Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2010). *Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features*. (CRESST Report 785). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Executive Summary

The purpose of this study is to examine the characteristics of reading test items that may differentially impede the performance of students with disabilities. By examining the relationship between select item features and performance, the study seeks to inform strategies for increasing the accessibility of reading assessments for individuals from this group. Including students with disabilities in large-scale, statewide assessment and accountability systems, as mandated by the Individuals with Disabilities Education Act (IDEA, 2004) and the "No Child Left Behind" (NCLB) Act of 2001 (NCLB, 2002), can help identify issues and guide instruction to improve education for these students.

Research on reading complexities for students has primarily focused on the role of vocabulary and sentence length, and has also touched upon issues of legibility such as format, typeface, and visuals. Although research reveals that readability measures are widely used and beneficial for matching students' reading levels with appropriate text, they do not identify the precise grammatical and cognitive components within sentences, paragraphs, or passages that may contribute to complexity for students with disabilities. While current research does address the critical need to accurately assess the reading performance of students with disabilities, a void in operationalizing reading complexity exists.

With the selected features in the present study (cognitive, grammatical, lexical, and textual/visual), we are building on previous research by exploring the role that these features may play on reading test items that may cause these items to function differentially for students with disabilities. Thus, the following research questions guided the analyses and reporting of this study:

1. How and to what extent does the cognitive complexity of reading assessments (item type, depth of knowledge, and scope) impact the performance of students with disabilities?

2. How and to what extent do textual/visual features of reading assessments (number of pages, words per page, typeface changes, point size changes, font changes, and unnecessary visuals) impact the performance of students with disabilities?

3. How and to what extent do lexical density (average lexical density and number of uncommon words) and lexical features (number of words greater than seven letters, number of relevant paragraphs, and number of words in items and relevant paragraphs) of reading assessments impact the performance of students with disabilities?

4. How and to what extent do grammatical features of reading assessments (subordinate clauses, complex verbs, passive voice, relative clauses, number of entities used as subjects, and noun phrases) impact the performance of students with disabilities?

5. Among the five major categories of complexity features in an assessment, which category or categories most discriminate in terms of reading performance between students with disabilities and their peers with no disabilities?

Method

To investigate our research questions, we evaluated current English language arts standardized assessments from three states to determine their cognitive, lexical, grammatical, and textual/visual complexity using differential item functioning (DIF) and discriminant analysis.

Population and Sample

The population for this study is students in grade 8 in three states. Because the states were not selected randomly, the level of generalizability of this sample to the population is limited and the results should be interpreted with caution when generalizing to the entire grade 8 student population.

Assessments

The present study analyzed a total of nine assessment forms from grade 8 reading assessments in three states, for a total of 490 reading test items. Active reading assessments and student data from three states were obtained with permission. The states are referred to as State A, B, and C to preserve anonymity. The State A assessments (from 2006 and 2008) consisted mostly of multiple-choice items, with a few extended-response items (usually only one per passage). These extended response items were not included in the present study because student data were not available for extended-response items. The State A assessments also contained field test items that were excluded from the analyses because student data were not available for these items. The State B assessments (from 2006, with four forms) consisted of multiple-choice items only. The State C assessments (from 2006, 2007, and 2008) were reading and writing assessments combined, which meant that some sections of the assessment consisted of a mixture of reading and writing items, while other sections solely assessed reading. Additionally, passages consisted of a blend of multiple-choice and extended-response items. Items that strictly measured writing standards were excluded from the present study; however, some items that have overlapping standards were retained.

Rating Guidelines Development

The development of a rubric used to evaluate test items and reading passages began with discussions about features that could interfere with the ability of students with disabilities to access content in reading assessments. A review of the literature and consultation with experts in the field resulted in our selection of five features that we used to capture accessibility for the purposes of this study:

1. Cognitive Features
2. Grammatical Features
3. Lexical A Features
4. Lexical Density B Features
5. Textual/Visual Features

The National Center on Educational Outcomes (NCEO) at the University of Minnesota provided "Considerations for Universally Designed Assessment Items" (Thompson, Johnstone, Anderson, & Miller, 2005). Based on literature reviewed in this report, and following consultations with experts in linguistics, we arrived at six grammatical features:

- passive verbs
- complex verbs (other than passive verbs)
- relative clauses
- subordinate clauses (other than relative clauses)
- complex noun phrases
- entities as subjects

Grammatical features were reduced to six to capture grammar usage efficiently. We also included a lexical features count, and other ways to rate clear format and clear visuals. The lexical features included counting the total number of words and the total number of unique words in order to compute lexical density. To capture difficult vocabulary, words consisting of seven letters or more were also counted (as adapted from Shaftel et al., 2006). Additionally, we used a corpus of common words (Bauman & Culligan, 1995) to count uncommon words and words of seven or more letters (lexical A features). Two categories, lexical A and lexical B, were created to distinguish features that are more or less likely to impact the construct being measured. Based on expert opinion and consensus of this research team, it was decided that changes to lexical A features may have less serious impact on the construct tested than lexical density B features.

Rating Process

Cognitive and grammatical categories were rated by external raters. Thirteen raters were assigned to one of two groups. The Grammar Group (7 raters) was responsible for rating the grammatical features. The Cognitive Group (6 raters) was responsible for rating the cognitive and textual/visual features. Raters from the applied linguistics department or with backgrounds teaching English as a Second Language were assigned to the Grammar Group, while all other raters were assigned to the Cognitive Group. A one-day training session was planned. Raters were presented with an overview of the features and instructions on the rubric, and then given released assessment items for practice. Items were first rated as a group, then discussed, then individually rated, and then discussed.

All passages, paragraphs, visuals, and items were assigned unique ID numbers to facilitate data entry and analyses. All paragraphs and visuals were numbered so that raters could list the relevant paragraphs and visuals that were necessary to answer an item. After numbering all passages 1 to 71, a random number generator was used to randomly distribute passages across the 6 raters for each group.

Results

To answer our five research questions, the reading assessments from three states were rated on 21 accessibility features in five general categories: (1) cognitive complexity, (2)

textual/visual complexity, (3) lexical A complexity, (4) lexical density B complexity, and (5) grammatical complexity. The cognitive complexity category included measures of passage and item types, depth of knowledge, and scope. The textual/visual complexity category included column count, number of pages, words per page, number of typeface changes, number of point size changes, number of font changes, and number of unnecessary visuals. The lexical A complexity category included a count of the number of words greater than seven letters in items and paragraphs, the number of relevant paragraphs, and the number of words in items and relevant paragraphs. The lexical density B complexity category included the average lexical density (total unique words per page/total words per page), and the number of uncommon words in items and relevant paragraphs. The grammatical complexity category included counts of the number of subordinate clauses, complex verbs, passive voice verbs, relative clauses, entities, and noun phrases.

Two different approaches were employed for analyzing the data: (1) a Multiple Discriminant (MD) approach and (2) a Differential Item Functioning (DIF) approach. In the MD approach, we examined the impact of the accessibility features between students with and those without disabilities across the entire test; and in the DIF approach, Differential Bundle Functioning (DBF) and Differential Test Functioning (DTF) approaches were applied to see the impact of the accessibility features on the entire test as well as on the individual test items or a group of test items (bundle of items) that share specific accessibility features.

Differential Level of Impact of Accessibility Features on Reading Assessments for Students with Disabilities: Results from a Multiple Discriminant Analyses

A multiple discriminant function provides a direct approach in comparing the impact of the 21 accessibility features on the performance of students with disabilities (SWDs) and non-SWDs. Data from the three states were used for this study. A data file was created in which a student's incorrect response (0 score) in each test item was replaced with ratings from each of the corresponding features. Therefore, the total score of a particular feature for each student was the incorrect responses (0) plus the rating of the feature for each individual item. As a result, 21 scores were created, with one for each accessibility feature. For example, using feature #2, item type, if a student responded to test item 1 incorrectly and item 1 had an item type rating of 4, then the student's incorrect score on item 1 would be 4. A similar procedure was used for creating other feature scores, thus the units of analysis in this study were individual students, not test items.

Results of the discriminant analyses suggest that: (1) some of the accessibility features, such as textual features, have more impact on reading than other features, and (2) some of these features have more differentiating powers between students with disabilities than others.

Results from Differential Item Functioning Using the Non-Compensatory Differential Item Functioning (NCDIF) Index and Logistic Regression

Results from discriminant analyses indicated that some of the accessibility features have more impact on student outcomes, particularly for students with disabilities. These results can be interpreted at the total test level. However, we also wanted to know whether some of the test items were more impacted because of these features than other items. We therefore conducted a series of Differential Item Functioning (DIF) and Differential Test Functioning (DTF) analyses. A multiple regression approach was then applied in order to examine the relationship between each of the complexity features and the signed uniform DIF ($\bar{\mu}_d$) findings. In our first set of analyses we examined the relationships of each individual feature to signed uniform DIF findings. Next we constructed a more comprehensive model that included measures from each of our complexity categories. These analyses were conducted at the item level across all nine reading assessments. The data were split into three strata representing items with low percentage range above guessing (PRAG) (0-11), moderate PRAG (12-29) and high PRAG (30 or above). As anticipated there was a strong correlation between item PRAG and the signed uniform DIF results (r = -0.762). The majority of items indicating DIF against SWDs had PRAG values over 30.

Of the 21 features modeled, 15 made significant contributions in the high PRAG items while only one feature had a significant r-square change within the both the low and moderate PRAG items. Each of the 15 significant features in the high PRAG items had model coefficients in the expected direction. The strongest individual cognitive feature was depth of knowledge. Among the grammatical features, complex verbs and subordinate clauses made the largest contributions. Lexical density at the passage level and words greater in length than seven letters that were present in items and their relevant paragraphs were each also strongly related to the DIF findings. Finally, a number of passage level textual/visual features were also significantly related to the DIF findings. Among those the strongest features were point size and font changes along with the number of unnecessary visuals.

We used a multivariate approach to examine whether unique contributions to DIF were present across the five complexity categories and multiple features. Latent variables were created: GRAMMAR (a combination of complex verbs and subordinate clauses), LEXICAL (lexical density at the passage level and words greater in length than seven letters that were present in items and their relevant paragraphs), TEXTVIS (unnecessary visuals, point size changes, and font changes), lexical_item (lexical density at the item stem level), the individual predictor depth of knowledge, and the individual predictor scope.

Among the six complexity variables examined, five had negative coefficients indicating increased DIF against SWD with higher values of the features. Therefore as the values of GRAMMAR, depth of knowledge, TEXTVIS, and lexical density (LEXICAL and lexical_item) increase, an item in the High PRAG category is more likely to exhibit DIF against SWDs. The scope variable has a positive coefficient; a result that is not consistent with

the rest of our findings. The TEXTVIS and scope measures were the two features that made the largest contribution toward explaining the variation in the DIF outcome.

Discussion

According to the literature cited in this report, students with disabilities perform substantially lower on standardized tests than students with no identified disabilities in both state (Abedi, Leon, & Mirocha, 2003; Altman, Thurlow, & Vang, 2009; Ysseldyke et al., 1998) and national assessments (Lee, Grigg, & Donahue, 2007). While part of this low performance may be explained by a student's specific disability or a student's lack of access to the general education curriculum, a major part of it may be attributed to the limitations of existing state assessments in addressing the needs of these students. That is, a part of the performance difference between students with disabilities and their peers without disabilities may be explained by accessibility issues. Current state assessments may not be sensitive enough to the needs and backgrounds of students with disabilities.

Based on the review of existing literature and consultations with experts in the field, we identified 21 accessibility features that could have major impact on the assessment outcomes of students with disabilities. These 21 features were grouped into the following five major categories: (1) cognitive features, (2) lexical A features, (3) lexical density B features (4) textual and visual features, and (5) grammatical features. The grouping of the 21 features into five main categories seems to be conceptually and analytically sound. Experts confirmed these categorizations, and results of factor analyses of the features within each category yielded strong evidence of internal consistency of the features within the five categories.

Two different analytical approaches were used in this study, a differential item functioning (DIF) approach and a discriminant analysis approach. In the DIF approach, using DTF and differential bundle functioning (DBF) methods, sets of test items representing a particular accessibility feature were compared across groups formed by students' disability status. Groups of accessibility features that behaved differentially across the two groups were identified and the level of impact on student reading performance was examined by their disability status in a multiple regression model. In the discriminant analyses model, the latent scores of five overall accessibility features were used as discriminating variables to identify the features that mostly discriminate the two groups.

Results of the two analyses consistently suggested that: (1) some of the accessibility features had more impact on reading than other features, and (2) some of these features had more differentiating powers than others between students with disabilities than students without disabilities.

Identifying features with the highest level of impact on the performance of SWDs has major implications for the assessment of these students, particularly when the features could be easily altered without changing the construct to be measured. There are many factors that affect student performance on assessments, and some of these are essential

components of the measures, such as the content and construct being measured. These factors cannot be altered because such changes might alter the construct being measured. However, some of these factors, such as textual/visual features, are incidental to the assessment and can be altered without having a major impact on the outcome of measurement. Another category, lexical A features, may provide an opportunity to reduce complexity without changing the construct being tested. For example, students with disabilities may find crowded test pages difficult and may experience fatigue and frustration when answering items in this format. Changing the test to include better readability for students does nothing to alter the construct, yet may significantly increase the performance of students with disabilities on such assessment items.

In summary, the results of this study can help the assessment community in two ways. First, by elaborating on some test accessibility features, this report may serve as a guideline for those who are involved in test development and the instruction and assessment of students with disabilities. Second, and more importantly, this report provides methodology for examining other features that may have a major impact on assessment outcomes for students with disabilities.

Table of Contents

Introduction

The study reported here investigated the characteristics of reading test items that impede the performance of students with disabilities. By examining the relationship between select item features and performance, the study sought to inform strategies for increasing the accessibility of reading assessments for individuals from this group. In the sections that follow, we summarize the literature that underlies our work and research questions; describe our methodology and present results and conclusions.

Including students with disabilities in large-scale, statewide assessment and accountability systems, as mandated by the Individuals with Disabilities Education Act (IDEA, 2004) and the "No Child Left Behind" (NCLB) Act of 2001 (NCLB, 2002), can help identify issues and guide instruction to improve education for these students. While this is the intent of accountability systems, including students with disabilities in assessments that have been developed for mainstream students may fail to present an accurate portrayal of their knowledge and skills. Some assessment characteristics may interfere with some students' abilities to access the content. Although students with significant cognitive disabilities participate in alternate assessments based on alternate achievement standards, most students with disabilities participate in general assessments. The policy of including these students in statewide assessments, therefore, assumes that the assessments are appropriate. For instance, according to data collected during the 2003-2004 school year, of the 48 reporting states and the District of Columbia, 41 states reported that at least 95% of students with disabilities participated in the statewide reading assessment (U.S. Government Accountability Office, 2005). In the 2006-2007 Annual Performance Reports, states reported that 82.8% of their middle school students with Individualized Education Plans (IEP) participated in regular reading assessments (Altman, Thurlow, & Vang, 2009). If regular assessments are not appropriate for students with disabilities, it becomes impossible to interpret and use results to make decisions about their education.

Students with disabilities have historically performed substantially lower on standardized tests than students with no identified disabilities (Abedi, Leon, & Mirocha, 2003; Altman et al., 2009; Ysseldyke et al., 1998). In the 2007 National Assessment of Educational Progress, also known as the *Nation's Report Card*, substantially fewer grade 8 public-school students with disabilities (34%) were performing at or above the basic level of reading than students without disabilities (76%) (Lee, Grigg, & Donahue, 2007). While the lower performance may be partly attributed to specific disabilities, or to lack of access to the general education curriculum, it may also be related to factors that impact accessibility. Therefore, it is critical to reduce such external factors on these assessments. Results from our previous studies indicated that some test items may well be affected by factors that result in differential functioning for students with disabilities in comparison to students without disabilities (Abedi, Leon, & Kao, 2008a, 2008b).

To increase participation of students with disabilities in general assessments, many states have allowed tests to be administered to students with accommodations, a common and allowable practice. Accommodations are changes to testing materials or the testing environment, such as changes in the presentation, setting, timing or scheduling, or response method (Thurlow, Elliott, & Ysseldyke, 2003). Accommodations are meant to increase the validity of test results for students with disabilities; however, they have been surrounded by challenges and controversies related to their administration and the interpretation of accommodated test results. In some cases accommodations may alter a test to the extent that accommodated and non-accommodated items are no longer comparable (Bielinski, Thurlow, Ysseldyke, Friedebach, & Friedebach, 2001). Consequently, there has been a shift toward Universal Design of Assessments. The underlying principle behind Universal Design is that the design of an assessment should be accessible to the largest number of students possible, thereby reducing the need for accommodations or other adaptations (Johnstone, 2003; Thompson, Thurlow, & Malouf, 2004). Universal Design of Assessments applies the principles of accessibility for all students when developing assessments to ensure that each student has a comparable opportunity to demonstrate achievement on the standards being tested (Dolan & Hall, 2001; Thompson, Johnstone, & Thurlow, 2002). Thompson et al. (2002, 2004) identified seven elements of universally designed assessments: (1) inclusive population; (2) precisely defined constructs; (3) accessible, non-biased items; (4) amenable to accommodations; (5) simple, clear, and intuitive instructions and procedures; (6) maximum readability and comprehensibility; and (7) maximum legibility. Experimental research findings show that students – including students with disabilities – scored significantly higher on a math assessment incorporating universal design principles (Johnstone, 2003). Moreover, students reported recognizing material better and found vocabulary and print more readable when it was presented using universal design principles.

The shift toward Universal Design reflects an increased interest in removing potential barriers to students' abilities to access assessment content. The present study is one in a series of research efforts conducted by the Partnership for Accessible Reading Assessments (PARA), part of a national effort to investigate accessibility principles to make reading assessments more accessible for students with disabilities (Thurlow et al., 2009). While there are many issues surrounding the assessment of students with disabilities, including the potential for over-, under-, and misidentification (Artiles, 2003; Koretz & Barton, 2003-2004), the present study focuses on aspects of reading assessments themselves, and assumes that students with disabilities were properly classified as such.

Who Are Students With Disabilities?

There are over 6.7 million children and youth (ages 3-21) with disabilities in the United States (U.S. Department of Education, 2007). Secondary school students (ages 12-17) with disabilities make up 43.7%, or 2.9 million, of the total population of students with disabilities (see www.ideadata.org for the most up-to-date information on numbers of

students with disabilities). Out of the total population of students with disabilities, nearly 39.2% (or over 2.6 million) were considered to have specific learning disabilities (SLD). Other cataloged disability categories include speech or language impairments (22.1% or nearly 1.5 million), health impairments (9.6%, or nearly six hundred fifty thousand), and mental retardation (7.6%, or just over half a million). Data are also collected on students with emotional disturbance, hearing impairment, orthopedic impairments, visual impairments including blindness, multiple disabilities, deaf-blindness, autism, traumatic brain injury, and developmental delays (U.S. Department of Education, 2007).

Literature Review

The present study focuses on reading assessments, and the role that identifiable features of those assessments may play for a specific subgroup of students – students with disabilities. This review of the literature looks at ways researchers have examined reading complexity in general and as it pertains to students with disabilities.

As mentioned earlier, assumptions that statewide assessments are appropriate for students with disabilities must be investigated before results can be used for decision-making. There is a void in examining item bias, test bias, and unnecessarily complex linguistic, lexical, and cognitive demands placed on this student population (Johnson, 2008; Koretz & Barton, 2003-2004). Additionally, in a recent study of a state standardized reading assessment, Kato, Moen, and Thurlow (2009) emphasized the benefit of examining items identified as functioning differently for students with disabilities. Identifying the cause of differential item functioning can aid in producing assessments that are accessible to all students. For example, the identification of specific cognitive, grammatical, and lexical features that reduce accessibility to assessments for students with disabilities can lead to improved test item writing and review, the goal of Universal Design of Assessment (Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008).

The following review of the literature first gives a brief overview of reading issues for students with disabilities; it then examines assessment features that may contribute to inaccessibility of reading assessments. The review provides a foundation for the assessment features investigated in the present study. Specifically, this study focused on five overarching areas, detailed in the Method section that may interfere with some students' abilities to access reading assessments: cognitive, grammatical, lexical (A and B), and textual/visual features.

Students with Disabilities and Reading

Reading is a complex task that involves multiple processes and abilities (Adams, 1990; Gough, Hoover, & Peterson, 1996; Siegel, 1993; Snow, Burns, & Griffin, 1998). It involves many perceptual, cognitive, and language comprehension processes, one of which is "the syntactic parsing of sentences" (Hall, White, & Guthrie, 1986, p. 90). In order to read, the reader uses a range of low-level processing skills, such as word recognition, and high-

order processing skills, such as accessing background knowledge, to make meaning of the text (National Accessible Reading Assessment Projects, 2006).

Reading problems pose one of the greatest barriers to success in school for students with learning disabilities (Swanson, 1999). A disability may affect the acquisition of reading skills that could impact reading development and comprehension in disparate ways. For example, students with specific learning disabilities, such as dyslexia, can exhibit weaker working memories (Shaywitz et al., 2003), difficulty with phonology (Shaywitz, 1998; Siegel, 1993; Snowling, Goulandris, & Defty, 1996), or less prior knowledge (Carr & Thompson, 1996).

In general, the most prominent challenges in reading skills for students with disabilities are basic print reading and reading comprehension (Gersten, Fuchs, Williams, & Baker, 2001). Webster (1994) suggested that syntactic knowledge (basic sentence comprehension) and syntactic awareness (the ability to reflect on sentence structure) are two important factors in reading comprehension. There is sufficient evidence to conclude that syntactic complexity causes difficulty in reading for students with learning disabilities (Kuder, 2008; Siegel & Ryan, 1984). For example, Siegel and Ryan (1984) confirmed through their research that reading disability has, as its root, a difficulty with phonological and syntactic skills. Children with reading disabilities are more challenged than their counterparts without disabilities with text that is syntactically irregular and complex in construction (Siegel & Ryan, 1984).

Cognitive Features

Much of the literature regarding cognitive features of assessments is derived from the literature on assessment alignment and validity. Researchers have proposed that cognitive processes be included when examining the validation of performance assessments. Including cognitive features when analyzing assessments allows researchers to consider the assessments' cognitive complexity, "thinking and reasoning activities elicited in assessment situations and the extent to which these activities are given preference in defining subject-matter achievement" (Baxter & Glaser, 1997, p. 1). Linn, Baker, and Dunbar (1991) added that for any analysis of assessments, "cognitive complexity of the tasks and the nature of the responses that they engender" (p. 19) must be considered. For the present study, the cognitive features we specifically examined were type of passage, item type (whether items were informational or inferential), depth of knowledge, and scope.

Type of passage. The type of passage refers to the overarching organizational structure of a passage; passage types are generally divided into two main categories – narrative and expository – but they also include description, persuasion, and poetry (Butler, Bailey, Stevens, Huang & Lord, 2004; Kamberelis, 1999; Paltridge, 2002). Research has shown that students with disabilities possess limited and delayed knowledge of narrative and expository text organization and structures, which could lead those students to experience difficulty comprehending, recalling, distinguishing between essential and

nonessential material, or organizing the material read (Cain, 1996; Englert & Thomas, 1987; Gersten et al., 2001; Hansen, 1978; Wong, 1980).

Inferences. Making inferences when reading is a complex cognitive skill that good readers accomplish in order to comprehend text. Inferences are necessary in reading comprehension because they link ideas and fill in details that are not directly mentioned in the text. Making inferences requires the reader to monitor what is read to establish coherence between different actions and events or to integrate the reader's general knowledge with the information provided in the text (Cain & Oakhill, 1999; Cain, Oakhill, & Bryant, 2004). Students with disabilities are often poor readers because they do not read strategically and monitor what is read or because they have difficulty drawing upon relevant background knowledge (Gersten et al., 2001; Williams, 1993); however, studies have shown that students with disabilities can be guided to improve their ability to draw inferences and thus, improve their comprehension (Gardill & Jitendra, 1999; Idol-Maestas, 1985).

Depth of knowledge. Based on the alignment rating criteria developed by Webb (1997, 1999), depth of knowledge ratings reflect the level of cognitive complexity of the information that students are expected to know. For assessments, depth of knowledge is related to: (a) the number of connections of concepts and ideas a student needs to make in order to produce a response; (b) the level of reasoning; and (c) the use of other self-monitoring processes. Webb created four levels used to judge depth of knowledge: (1) recall, (2) skill/concept, (3) strategic thinking, and (4) extended thinking. Ratings include the following cognitive processes: the number of ideas integrated, depth of reasoning required, knowledge transferred to new situations, multiple forms of representation employed, and mental effort sustained. Items rated with a score of 1 reflected a low level of cognitive complexity whereas items rated with a score of 4 reflected a high level of cognitive complexity in which the items were expected to elicit a deep level of knowledge from students. There is evidence of a moderate to strong relationship between depth of knowledge ratings and item complexity, item difficulty, and student performance. Items judged to require high depth of knowledge were often the same items that were also rated most complex and items that students found most difficult (Herman, Webb, & Zuniga, 2003). Webb's depth of knowledge criteria were initially used to align science and mathematics standards and assessments, but they have also been adapted for use on language arts assessments (Webb, Horton, & O'Neal, 2002; Wixson, Dutro, & McDaniel, 2002).

Scope. Frederiksen and Collins (1989) proposed that scope – or the range of knowledge, skills, and strategies required to do well on the assessment activity – must be included in the standards used in designing testing systems to foster the development of the cognitive traits that the tests are designed to measure. In examining the language use found in language tests, Bachman and Palmer (1996) narrowed the idea of scope to the relationship between input and response and described scope as "the amount or range of input that must be processed in order for the test taker or language user to respond as expected" (p. 55). They further divided "scope of the relationship" into broad and narrow

scope. Broad scope assessment tasks are "main idea" reading comprehension questions that relate to the content of an entire passage. Narrow scope assessment tasks require processing only a limited amount of input, such as reading comprehension questions that center on specific details or limited parts of the reading passage. In their analyses of the language demands in English language proficiency tests with respect to scope, Wolf et al. (2008), using a 4-point scope rating scale, found that the items based on reading passages were of narrow scope that only asked students to recall literal information.

Grammatical Complexity

Rimmer (2006) stated that although grammar competence can be tested as an independent component of reading, grammar complexity is not a clearly defined construct. Other studies have confirmed the value of examining the general concept of grammatical complexity. Hall, White, and Guthrie (1986) suggested the need for investigating which category of words–verbs, conjunctions, nouns, etc.–should be examined as contributors to the abundant complex syntax found in schools' written texts and tests. Wang (1970), in an experimental study examining seventy-five sentences with varying syntactic structures, tentatively concluded that surface structure complexity plays an important role in sentence comprehension.

Research has addressed specific linguistic features contributing to grammatical complexity in reading content. Klare's (1974-75) comprehensive review of the literature revealed 12 published formulas or similar devices for predicting readability that focused on particular grammatical structures rather than on word, syllable, or sentence counts. These formulas considered personal reference words, pronouns, connectives, prepositional phrases, nouns, complex noun phrases, and select subordinate structures. Zacks, Speer, and Reynolds (2009) suggested that reading comprehension in narrative texts may be influenced by conceptual cues such as changes in characters in a story. These changes may signal to a reader that their existing mental model is no longer adequate. The linguistic feature of "entities as subjects" in the present study is intended to capture the number of different or new characters introduced by paragraph into narrative passages.

Hess and Biggam (2004) stated that the difficulty of text passages is a key factor in reading comprehension. They delineated a progression of complexity: sentence structure develops in increasing complexity from simple noun and verb (grade 1) to inclusion of causal phrases (grade 2), passive voice, abstract or descriptive language (grades 3-4), dialect and other linguistic variants (grades 5-6), culminating in a wide use of uncommon words and varied sentence structure (grades 7-8 and high school). Using a similar curriculum continuum, Hess (2008) outlined the grades in which increasingly complex sentence structures can be used on assessments. These structures range from simple sentences (grade 3) to simple and compound sentences (grade 4) including phrases and clauses (grade 5), and continuing with the expectation of varied sentence structures across the remaining grades (grades 6-12).

Measuring grammatical complexity. Studies have been conducted suggesting the use of broad measures of grammatical complexity. For example, Chall, Bissex, Conard, and Harris-Sharples (1996) presented a holistic method of assigning text difficulty levels using five or six broad characteristics (lexical features, sentence length and complexity, cognitive complexity, and idea density and difficulty) to assign a readability level from one to sixteen. These levels served to match individual students with text at their reading ability. This practical guide was developed as a quick qualitative means of ascribing readability levels by text writers, librarians, teachers, and others interested in reading comprehension. It did not precisely define or count language features that may contribute to complexity.

Bailey, Butler, and Sato (2005) established a rubric to rate language demands for students to meet standards in reading and science. Language demands were divided into two categories: academic language demands and linguistic skills. The latter category included the skill "phrases and sentences" which carried the goal to "determine meaning of spoken and written phrases and sentences" (p. 10). "Phrases and sentences" were categorized by complexity: low, medium, or high, but did not address specific grammatical structures that may contribute to complexity.

Many studies have focused on grammar in math assessments. An experimental study by Larsen, Parker, and Trenholme (1978) broadly identified linguistic complexity in terms of the presence of dependent clauses in math word problems. In another study, specific language characteristics were considered as sources of construct-irrelevant variance on a math exam for 4th, 7th, and 10th graders. Language characteristics had from small to medium effects on item difficulty at different grade levels for all examinees regardless of language proficiency or disability. Features shared with the present study included passive voice, clauses, complex verbs, conditionals, relative pronouns, pronouns, and prepositions (Haladyna & Downing, 2004). Another in-depth review of the impact of multiple language features on standardized math assessments included passive voice, clauses, complex verbs, infinitives, pronouns, conditionals, and prepositions. Interestingly, no outcome differences among English learners, students with disabilities, and other students were found (Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). The authors concluded that the testing company's rigorous test and item development procedures could be credited for these unbiased results. Lord (2002) reviewed multiple math assessments for the impact of grammatical features including subordinate clauses on students' performance. These studies examined correct response rates to NAEP math questions and to released and linguistically modified standardized test questions. Lord concluded that incorrect responses on items with subordinate clauses accounted for ten percent of the variance in NAEP scores. Additionally, students in lower level math classes benefited from item modifications that reduced linguistic complexity in general, and specifically from eliminating subordinate clause2.

Previous research has addressed methods of assigning text readability or complexity ratings (e.g., Botel, 1972; Chall & Dale, 1995; Dale & Chall, 1948; Klare, 1974-1975; Stenner, Burdick, Sanford, & Burdick, 2006). Some researchers have developed word

and structure frequency counts to capture grammatical complexity (Klare, 1974-75). Word count programs are more common than sentence structure counts, possibly due to the labor intensiveness of the latter. Choosing the structures to count, training raters to recognize the selected structures and the subsequent lack of rater reliability are reasons why sentence structure counts are less common (Roland, Dick, & Elman, 2007).

Grammatical complexity and students with disabilities. Another body of literature specifically addressed the impact of linguistic complexity on students with disabilities. Some of these linguistic features discussed below are captured by the rubric used in the present study. Research suggests that all children acquire linguistic rules in the same order; however, children with language development problems acquire the rules more slowly and at a later age than children without such problems (Alvermann, 1981; Tyack & Gottsleben, 1977). Venable (2003) gave examples of several linguistic features that can cause difficulty for readers with learning disabilities. She cited the use of a relative clause with the omitted pronoun "that" (null pronoun) and noun phrases that consist of two nouns (e.g., dirt floors) as two frequently occurring sources of sentence complexity.

In an experimental study of 24 English-speaking adults with mild and borderline learning disabilities, over half of the participants were able to demonstrate understanding of six constructions labeled as easy (e.g., noun-shoe, verb-eating). Only 11 percent of the participants were able to exhibit comprehension of sentences labeled as complex, such as those that used passive voice and relative clauses (Jones, Lord, & Finlay, 2006).

Chappell (1985) discussed language disabilities using Guilford's (1967) Structure of the Intellect model. Examples of comprehension difficulties experienced by students with language disabilities occurred in: (1) passive structure; (2) complexly embedded sentences with interrelationships (as found in relative clauses); and (3) modal auxiliaries (e.g., could, would, most, might, or should).

Tyack and Gottsleben (1977) established procedures for listening to and analyzing the speech of children with serious language delays. Their analysis shares the following linguistic features with the present study: word count, modals, verb tense, prepositions, and relative and subordinate clause counts. Although the Tyack and Gottsleben study focused on oral language, it is difficult to comprehend in reading or listening structures that a student cannot or has not produced orally (Gennari & MacDonald, 2008).

In an experimental study of students with disabilities, Abrahamsen and Shelton (1989) administered a social studies reading passage with 10 short answer questions to 4 groups of 23 students each. One passage was presented in the original version as found in the text, one passage was modified syntactically, one was modified semantically, and one was modified both syntactically and semantically. Syntactic modifications consisted of changing passive to active voice, changing past perfect tense to simple past tense, and eliminating relative clauses. Student groups assigned to reading passages with syntactic modifications scored significantly higher than those in the control group or in the group with only semantic modifications. The authors concluded that modified passages resulted in improved performance for students with disabilities.

Lexical Complexity

The two most consistent identifiers of text complexity are vocabulary and sentence length. According to Chall and Dale (1995), "Vocabulary difficulty is a strong predictor of text difficulty. Knowledge of words has been a strong measure of child language development, of reading comprehension, and of verbal intelligence" (p. 82). These identifiers are therefore the two principal components used in many readability measures.

Readability measures have a history dating from the early 1920s. Klare (1974-1975) reviewed approximately 31 frequently used readability formulas, which included measures that count language variables to produce indices of probable difficulty. The formula computations were made up of 58 different measures of sentence length, syllable, and word counts. For example, 16 formulas included the average sentence length in words; six included the number of one syllable words per 100 words, two counted the average words per independent clause, and one counted word length in syllables. Early ratings were used to rank reading passages in order of complexity and later to assign a grade-level equivalency. The term "readability" has recently been replaced by the word "leveling" to rate reading passages in the early grades.

The role of vocabulary in text difficulty forms the basis of the well-established Classic Readability Theory (Chall & Dale, 1995). The Lexile Theory, developed from the Classic Readability Theory, also emphasizes the two components, word familiarity, and sentence complexity, as predictors of overall text complexity.

Another methodology supporting a determination of text readability or complexity is the Lexile Framework for Reading (Stenner et al., 2006). The two key components of this framework are word familiarity based on frequency of occurrence in a 550-million word corpus; and sentence length, a proxy for syntactic complexity. The resulting lexile (L) scale ranges typically from 200L for beginning readers and texts to 1700L for advanced readers (see North Carolina Public Schools Web site at www.ncpublicschools.org/accountability/parents/lexiles/educators?&print=true). That does not necessarily result in a fine grained analysis. A sentence with several coordinate or main clauses is not necessarily as complex as one with several subordinate clauses. These two sentences can be the same length but have different degrees of complexity; the sentence with coordinate clauses is likely to be less complex than one with several subordinate clauses.

Textual and Visual Features

Font. To facilitate accurate coding of these features, the frequently used term "typeface" (often used interchangeably with "font") was subdivided into several features using common Microsoft Word terminology. We used "font" to refer to the names of screen or print text such as Times New Roman, Courier, or Helvetica. Text that is presented in bold, italics, or underlined (Microsoft Word "font styles") and text that is all caps, outline, or shadow (Microsoft Word "effects") were incorporated under the feature

of "typeface" for the purposes of this report. Point size refers to height of letters. Our study does not attempt to determine optimal font, typeface, and point size, but rather to examine the relationship between changes in these features and performance on standardized reading tests.

Research indicates that choices in font and typeface may influence print legibility, (Mansfield, Legge, & Bane, 1996; Roethlein, 1912; Tinker, 1963). Legibility is most accurately measured by speed of reading (Bloodsworth, 1993; Tinker, 1963). Mansfield et al. studied the legibility of various fonts and typefaces on readers with normal and low vision. They found that proportionally spaced fonts such as Times-Roman, with letters taking up different amounts of horizontal space; and fixed width fonts such as Courier-Bold, with each letter allotted the same amount of horizontal space, affect three reading performance areas: (1) maximum reading speed (the speed that is not impacted by print size); (2) reading acuity (the smallest size at which print can be read); and (3) critical print size (the smallest print size read at the fastest speed). Mansfield et al. (1996) concluded that some categories of low vision readers are more impacted by font features. Specifically they stated that, "If print size is smaller than the critical print size, the choice of font could make a functionally significant difference in reading speed and accuracy" (p. 1500).

Typeface. Text presented in bold, italics, underlined, or all caps is included under the category of typeface for the purposes of this report. Typeface is listed as a legibility feature that needs to be considered in assessments under Universal Design principles. Thompson, Johnstone, and Thurlow (2002) stated that the degree of legibility of text found in assessments can bias results if items contain physical features that interfere with a student's focus or understanding of the construct an item is intended to assess. Characteristics of legible typeface include the use of boldface instead of italics for more visibility and the avoidance of text printed in all caps (Roethlein, 1912; Thompson et al., 2002).

Line length. Thompson et al. (2002) suggested that line length (or, the length of the line of text between the left and the right margin) was a factor for legibility when considering Universal Design of Assessments. While research on line length in relation to reading comprehension has not been conclusive, researchers have made recommendations on optimal line lengths based on reading rates and analyses of eye movements (Tinker, 1963). An optimal line length has been estimated at 52 characters per line (Rayner & Pollatsek, 1989), not to exceed 70 characters (Spencer, 1968, as cited in Dyson & Haselgrove, 2001). Another study recommended 24 picas, or about 4 inches (Worden, 1991, as cited by Thompson et al., 2002). Dyson and Haselgrove (2001), in their study of line length on computer screens, found that a medium length of 55 characters was effective for reading at normal and fast speeds as well as for comprehension.

Visual features. Visual features include pictures, illustrations, tables, and graphs that accompany text; and the usefulness of these visual features should be considered when integrated in assessments. Pictures and illustrations, as summarized by Shorrocks-Taylor

and Hargreaves (1999), can be categorized into three types: (1) decorative – not related to the question and serve no instructional purpose; (2) related – have the same context as the questions and are used to support text and emphasize ideas; and (3) essential – not repeated in text, but the text refers to them and they have to be read or worked with to answer the question. In a review of research that examined the effects of pictures on reading accuracy and comprehension, Filippatou and Pumfry (1996) found that pictures are not uniformly effective in all prose-reading situations and not all types of pictures are equally effective for children with differing reading abilities. Moreover, the authors found that (1) picture effects are expected to vary as a function of relevant student characteristics, and the value of illustrations appears somewhat questionable for students with learning disabilities and poor readers, (2) the degree of picture facilitation expected depends on relationship between particular learning task and kind of pictures provided (e.g., detailed, text-redundant, relational pictures seem to help some readers comprehend more), (3) the facilitative effect of pictures on inferential comprehension depends on type of passage (mainly abstract passages), (4) combined strategy of mental imagery and text pictures are more facilitative in reading comprehension than text pictures only, and (5) the facilitative effects of illustrations depend on type of pictures in relation to type of learning material (e.g., in science and social studies expository text). With respect to Universal Design, Thompson et al. (2002) contended that the inclusion of visual features on assessments should consider the needs of some students in which visual-related challenges may be present. Even without visual challenges, students:

> may be unnecessarily distracted due to an inability to shift their focus between the relevant information and extraneous or irrelevant information. For example, illustrations added for interest may draw attention of some students away from the construct an item is intended to assess. (p. 18)

Overall, researchers agree that the usefulness of visual features depends on the student's reading abilities and learning styles (Filippatou & Pumfry, 1996; Shorrocks-Taylor & Hargreaves, 1999; Thompson et al., 2002).

Research Questions

Research on reading complexities for students has primarily focused on the role of vocabulary and sentence length, and has also touched upon issues of legibility such as format, typeface, and visuals. Although research reveals that readability measures are widely used and beneficial for matching students' reading levels with appropriate text, they do not identify the precise grammatical and cognitive components within sentences, paragraphs, or passages that may contribute to complexity for students with disabilities. While current research does address the critical need to accurately assess the reading performance of students with disabilities, a void in operationalizing reading complexity exists.

With the selected categories in the present study (cognitive, grammatical, lexical, and

textual/visual), we are building on previous research by exploring the role that these features may play on reading test items that would cause them to function differentially for students with disabilities. Accounting for all of the factors and features that may interfere with students with disabilities' abilities to access test content may be difficult, because the data may be too large and too intricately intertwined to quantify or measure. However, by narrowing our focus to these features, we hope to shed light on accessibility issues in the hopes of improving assessment and accountability for students with disabilities.

Thus, the following research questions guided the analyses and reporting of this study:

1. How and to what extent does the cognitive complexity of reading assessments (item type, depth of knowledge, and scope) impact the performance of students with disabilities?

2. How and to what extent do text features of reading assessments (number of pages, words per page, typeface changes, point size changes, font changes, and unnecessary visuals) impact the performance of students with disabilities?

3. How and to what extent do lexical density (average lexical density and number of uncommon words) and lexical features (number of words greater than 7 letters, number of relevant paragraphs, and number of words in items and relevant paragraphs) of reading assessments impact the performance of students with disabilities?

4. How and to what extent do grammatical features of reading assessments (subordinate clauses, complex verbs, passive voice, relative clauses, number of entities, and noun phrases) impact the performance of students with disabilities?

5. Among the five major categories of complexity features in the assessments, which category or categories most discriminate between students with disabilities and their peers with no disabilities in terms of reading performance?

Method

To investigate our research questions, we conducted this study in two phases. In phase I we evaluated nine English language arts standardized assessments from three states to determine their cognitive, lexical, grammatical, and textual/visual complexity using the rating guidelines at Appendix A and the rating coding forms at Appendix B. The population and samples, assessments, student data, rating guidelines, raters and ratings, and complexity features are detailed below. In phase II, we compared the results of our complexity ratings and student assessment scores to determine the features that had the greatest impact on students with disabilities using a multiple discriminant analysis and a differential item functioning analysis employing the NCDIF index and logistical regression.

Population and Sample

The population for this study is students in grade 8 in three states. Because the states were not selected randomly the level of generalizability of this sample to the population is limited and the results should be interpreted with caution when generalizing to the entire grade 8 student population.

Assessments

The present study analyzes a total of nine assessment forms from grade 8 reading assessments in three states, for a total of 490 reading test items. Active reading assessments and student data at the item level from three states were obtained with permission. The states are referred to as States A, B, and C to preserve anonymity. States were chosen based on their willingness to volunteer for the study. The State A assessments (from 2006 and 2008) consisted mostly of multiple-choice items, with a few extended-response items (usually only one per passage), that were not included in the present study because student data were not available for extended-response items. The State A assessments also contained field test items that were excluded from the analyses because student data were not available for these items. The State B assessments (from 2006, with four forms) consisted of multiple-choice items only. The State C assessments (from 2006, 2007, and 2008) were reading and writing assessments combined, which meant that some sections of the assessment consisted of a mixture of reading and writing items, while other sections were solely reading. Additionally, passages consisted of a blend of multiple-choice and extended-response items. Items that strictly measured writing standards were excluded from the present study; however, some items that had overlapping standards were retained. Additionally, some passages were repeated across years with both minor changes and new items.

Data

Student data from the corresponding years were obtained. To avoid potentially confounding issues, English language learners (ELLs) were removed from each sample. For our DIF analysis (described in detail later), we examined two focal groups of interest. These two groups included all students with disabilities and students identified as having a specific learning disability (SLD), one of the 13 disability categories for special education services. Information on disability type was not available in State A, thus our analyses for students with SLD were confined to the seven assessments from States B and C. The reference group for each analysis consisted of students without disabilities who were not ELLs. As expected, there was a large, unbalanced population size between the focal and reference groups. Students without disabilities far outnumbered the two focal groups in all assessments. When the population of students without disabilities was more than nine times larger than the focal group of interest we selected a random sample of students without disabilities that was roughly nine times larger than each focal group of interest. This was done to remain within the conditions of the simulation study upon which our DIF method is based (Leon, 2009).

Population sizes and raw score means are presented in Tables 1 and 2 for the analyses for all students with disabilities, and for the students with SLD, respectively. Students without disabilities in States A and B tended to score about 1 standard deviation higher than students with disabilities. In State C, this difference was somewhat larger. The populations of students with an SLD were about one half as large as the populations of all students with disabilities. All population group sizes were sufficient to apply the DIF and the discriminant analyses techniques used in this study.

Table 1. Reading Assessments and Sample Sizes for Analyses of All Students with Disabilities by State Assessment Form

State	Year	Assessment	Number of items	Students with Disabilities		Students without Disabilities		
				n	Mean Raw Score	n	Mean Raw Score	Total Standard Deviation
A	2006	Grade 8	48	2814	18.98	24908	31.90	10.76
A	2008	Grade 8	48	2456	20.54	22333	32.91	10.77
B	2006	Grade 8 Form 1	56	2294	27.88	20578	38.46	9.46
B	2006	Grade 8 Form 2	56	2527	28.73	21208	38.75	9.70
B	2006	Grade 8 Form 3	56	2290	28.48	20563	38.22	8.90
B	2006	Grade 8 Form 4	56	2361	28.43	21174	38.41	9.16
C	2006	Grade 8	58	4686	27.04	42108	39.74	9.66
C	2007	Grade 8	56	4710	25.18	42478	37.16	9.57
C	2008	Grade 8	56	4674	26.61	41861	39.21	9.59

Note. Students without disabilities refers to the reference group created as a random sampling of students without disabilities not exceeding nine times greater than the focal group.

Table 2. Reading Assessments and Sample Sizes for Analyses of Students with Specific Learning Disabilities by State Assessment Form

State	Year	Assessment	Number of Items	Students with Specific Learning Disabilities		Students without Disabilities		
				n	Mean Raw Score	n	Mean Raw Score	Total Standard Deviation
B	2006	Grade 8 Form 1	56	1180	29.01	10631	38.55	9.32
B	2006	Grade 8 Form 2	56	1413	29.55	12685	38.75	9.62
B	2006	Grade 8 Form 3	56	1210	29.64	10849	38.23	8.71
B	2006	Grade 8 Form 4	56	1219	29.65	10951	38.32	8.99
C	2006	Grade 8	58	2383	25.95	21053	39.77	9.77
C	2007	Grade 8	56	2327	24.77	20973	37.18	9.55
C	2008	Grade 8	56	2411	25.78	21765	39.21	9.60

Note. Disability type information was not available for State A. Students without disabilities refers to the reference group created as a random sampling of students without disabilities not exceeding nine times greater than the focal group.

Rating Guidelines Development

The development of the rating guidelines used to evaluate test items and reading passages began with discussions about features that could interfere with the ability of students with disabilities to access content in reading assessments. We arrived at five areas to capture accessibility for the purposes of this study:

- Cognitive Features
- Grammatical Features
- Lexical A Features
- Lexical Density B Features
- Textual/Visual Features

Table 3 summarizes the features examined in this study. The major cognitive areas included passage type, item type, depth of knowledge (Webb 1997, 1999), and scope (Bachman & Palmer, 1996; Wolf et al., 2008). Although Webb's depth of knowledge scale was developed to conduct alignment between content and standards, we adapted it for use in our study. We also reduced the scale from 4 points to 3, determining that a Level 4 (Extended Thinking) would not be applicable to multiple-choice type items. For scope, we considered, on a 5-point scale, the extent to which students would need to refer to the reading passage to answer an item. While Bachman and Palmer (1996) and Wolf et al. (2008) used a 4-point scale to measure "scope of the relationship," we expanded our scale

to five points after examining a variety of reading passages and determining that five points would be more nuanced.

In addition to cognitive features, we considered linguistic features. The National Center on Educational Outcomes (NCEO) at the University of Minnesota provided "Considerations for Universally Designed Assessment Items" (Thompson, Johnstone, Anderson, & Miller, 2005) which included the following points:

- item tests its intended construct
- item respects the diversity of the assessment population
- item has concise and readable text
- item has a clear format for text
- item has clear visuals
- item allows changes to format without changing meaning or difficulty

Following these considerations, we focused the present study on the third, fourth, and fifth bullet points. For "concise and readable text," we operationalized it to measure linguistic features (which refer to both grammatical and lexical features). A 17-item linguistic complexity checklist by Shaftel et al. (2006) served as a starting point. Based on literature reviewed earlier in this report, and following consultations with experts in linguistics, we arrived at six grammatical features:

- passive voice
- complex verbs (other than passive verbs)
- relative clauses
- subordinate clauses (other than relative clauses)
- complex noun phrases
- entities as subjects

Grammatical features were reduced to six to capture grammar usage efficiently. We also included a lexical features count, and other ways to rate clear format and clear visuals. The lexical features included counting the total number of words and the total number of unique words in order to compute lexical density (lexical B features). To capture difficult vocabulary, words consisting of seven letters or more were also counted (as adapted from Shaftel et al., 2006). Additionally, we used a corpus of common words (Bauman & Culligan, 1995) to count uncommon words and words of seven or more letters (lexical A features). Two categories, lexical A and lexical B, were created to distinguish features that are more or less likely to impact the construct of an assessment. Based on expert opinion and consensus of this research team, it was decided that changes to lexical A features may have less serious impact on a test's construct than lexical density B features. Future experimental research can lend support to the idea of two distinct lexical categories through field tests. By testing a non-student with disability (SWD) sample of students on both lexical A and reduced complexity lexical A features, the accuracy of this categorization can be tested.

Cognitive and grammatical features and some of the textual/visual features required the use of external raters, described below. Lexical A and lexical density B features were counted through the use of computer programs. Counts were created at both the passage level and the paragraph level, so that averages could be taken across different paragraphs if needed. Lexical counts were also taken at the item level, including combinations with the entire passage with each individual corresponding item.

Of important note is the assessment component (i.e., passage, paragraph, or test item) to which the complexity rating was applied (see Table 3). When examining accessibility features, the complexity of reading assessment demands were examined at both the passage and the test item levels, together with their relationships to each other. For the purposes of this study, *passage* refers to the entire text of the reading passage, including the directions or introduction to the passage, but excluding the items. The use of the term *passage* in this report also refers to poems, plays, narrative fiction, or any presentation of text that has corresponding items to solve. Passages were broken down into *paragraphs* using natural paragraph breaks in most cases, with some exceptions for poems and narrative fiction. Some poems consisted of one paragraph only. Other poems were broken down by stanza, if there were natural stanzas as printed in the assessment. Some works of narrative fiction consisted of mostly dialogue, in which case some dialogue was grouped with subsequent natural paragraphs as one paragraph. Paragraphs in plays usually consisted of breaks in character, and may have included stage directions with a character's dialogue if applicable. All passages were pre-numbered with paragraphs before distribution to raters. *Items* in this study refer to the multiple-choice questions in the assessments.

Paragraph-level ratings were conducted on some features such as grammatical and both lexical categories in order to explore the relationship between an item and the corresponding paragraphs in which the answers could be found. Ratings were also provided on the specific paragraphs and/or visuals where answers to the items could be found, or where students needed, at a minimum, to read and look for the answer. For example, if the answer was found in paragraphs 2 and 3 and visual 1, ratings would indicate paragraphs 2, 3, and visual 1. If the answer required inference across an entire passage, all paragraphs and visuals for the passage were listed. *Visual* in this study refers to any table, chart, graph, picture, or illustration printed in the assessment with a corresponding passage. For more specific information on the rating guidelines, see Appendix A.

Table 3. Summary of Features Measured

Feature	Description/Scale	Assessment Component
Cognitive		
Passage type[a]	Descriptive Narrative Expository Poetry Persuasive	Passage level
Item type	Informational Inferential	Item level
Depth of knowledge	Level 1: Recall Level 2: Skill/Concept Level 3: Strategic thinking	Item level
Scope	Scale of 1 (item can be answered without referring to the passage) to 5 (student needs to understand the entire passage)	Item level
Answer location[b]	Paragraph numbers and/or visuals needed to answer an item	Item level
Grammatical		Paragraph, passage, and item level (except for entities)
Passive voice	Ordinal	
Complex verbs[c]	Ordinal	
Relative clauses	Ordinal	
Subordinate clauses[d]	Ordinal	
Complex noun phrases	Ordinal	
Entities as subjects	Ordinal	
Lexical A		
Total number of words	Ordinal	Relevant parag. and item levels
Words with 7+ letters	Ordinal	Relevant parag. and item levels
Paragraph s	Ordinal	Number of relevant paragraphs
Lexical Density B		
% of unique/total words	Ordinal	Relevant parag. and item levels
Uncommon words	Ordinal	Relevant parag. and item levels
Textual/Visual		
Columns	1 or 2, as a proxy for line length (assumes 2 column-format would be roughly 4 inches in length, while 1 column formats would be wider across the page)	Passage level
Words per page	Ordinal	Passage level
Number of pages	Conversion of inches to decimal to measure the number of pages a passage spans in order to calculate blank space	Passage level
Typeface	Measure of change (bold, italic, underline)	Paragraph and item level
Font	Measure of change (font type)	Paragraph and item level
Point size	Measure of change	Paragraph and item level
Visual	Visual type (table, chart/graph, other) Necessary (yes, no)	Passage level

Table 3. Summary of Features Measured (continued)

Notes: [a] There was little variation among passage types. Therefore this feature was not included in the analyses and was dropped from further references as a feature.

[b] This entry was identified by the cognitive raters in order to capture relevant paragraph information for other features. "Answer location" is not an individual accessibility feature.

[c] Complex verbs other than passive voice

[d] Subordinate clauses other than relative clauses

Rating Process

Rater recruitment. Cognitive and grammatical categories were rated by external raters. An advertisement was sent out to the departments of applied linguistics and education at UCLA recruiting those with backgrounds in linguistics, education, test development, and/or teaching. Some raters were also recruited through word of mouth. Thirteen potential raters were invited to attend the first training session. Raters were assigned to one of two groups. The Grammar Group (7 raters) was responsible for rating the grammatical features. The Cognitive Group (6 raters) was responsible for rating the cognitive and textual/visual features. Raters from the applied linguistics department or with backgrounds teaching English as a Second Language were assigned to the Grammar Group, while all other raters were assigned to the Cognitive Group.

Rater training. Initially, a one-day training session was planned. Raters were introduced to the study and then separated into their respective groups. Raters were presented with an overview of the features and instructions on the rating guidelines, and then given released assessment items for practice. Items were first rated as a group, then discussed, then individually rated, and then discussed. It became clear at the end of the day that more training would be needed. The raters were given more released items to practice at home and another meeting was scheduled for both groups for the following week. Discussions as a group during the meetings helped reinforce the rating guidelines scale definitions, especially for the Cognitive Group. For the Cognitive Group, the second meeting seemed sufficient to move forward with rating.

For the Grammar Group, however, more training was needed, specifically with practice in learning to recognize the grammatical features. Some raters had stronger prior knowledge of grammar than others. Also, specific issues came up with the definition of certain grammatical features, and modifications were made to the rating guidelines following discussions with the raters. For example, reduced relative clauses and reduced passive verbs were not initially counted in the original rating guidelines, and were subsequently added. Complex noun phrases were also modified from the original rating guidelines to remove the simpler 1 Noun + 1 prepositional phrase (e.g., "the history of chocolate") and instead, to count 1 Noun + 2 prepositional phrases or 1 adjective + 1 Noun + 1 prepositional phrase (see Appendix A for more detail). These changes,

however, led to some confusion for some of the raters, and some required one-on-one meetings for clarification. Raters in the Grammar Group were given more passages and items to rate on their own as practice. It became clear that some raters were ready to move on to the actual rating (Raters 3, 4, 5), while other raters needed to be given more clarifications on the rating guidelines (Raters 1 and 2 did not count noun phrases or entities correctly), while still others failed to grasp some grammatical features altogether (Raters 6 and 7 did not fully understand passive and complex verbs or clauses). Some one-on-one training was given to Raters 1, 2, 6, and 7. Finally, a decision was made to drop Rater 7 after the additional training failed to produce results. The other raters were retained for a final group of six.

Assigning ID numbers to passages and paragraphs. All passages, paragraphs, visuals, and items were assigned unique ID numbers to facilitate data entry and analyses. All paragraphs and visuals were numbered so that raters could list the relevant paragraphs and visuals that were necessary to answer an item. Also, for logistical purposes only, to facilitate random distribution of passages to raters, all passages were numbered 1 to 71, unrelated to their ID numbers. However, some "passages" were not traditional passages, for example, a chart with two items. Thus, the few "loose" items were grouped together as one passage. Also, passages that were repeated across years, that is, once in 2006 and again in 2008, were assigned the same passage number and distributed to the same rater. After numbering all passages 1 to 71, a random number generator was used to randomly distribute passages across the 6 raters for each group.

Rater reliability and final rating process. Rater reliability results guided the final rating process and some of the distribution of reading passages to individual raters. In order to gauge reliability, traditionally, 25% of the items are rated and a reliability score is then computed. Since it was easier logistically to distribute 25% of the passages (rather than items), 18 out of the 71 total passages were randomly selected and distributed among the six raters in each group. Each rater was given six of the passages so that every passage was individually rated by a total of two raters. The 18 passages contained 119 out of the 490 total items.

For the Cognitive Group, the specific features of interest for reliability were: item type, scope, and depth of knowledge. Reliability was computed using Cronbach's alpha. Table 4 below displays the reliability results for the items from the first 18 passages:

Table 4. Reliability Results for the Cognitive Group for Items from 18 Passages

Feature	Overall Reliability	Removing Rater 2	Removing Raters 2,6
Item type	0.77	0.78	0.83
Scope	0.68	0.78	0.80
Depth of knowledge	0.63	0.73	0.79

Initial results indicated that reliability coefficients for each of the three features were below ideal. However, reliability improved for each of the three features after removing the ratings from Rater 2, and even more so after removing Rater 6's ratings as well. Initial analyses of ratings indicated that these two raters were not consistent with other raters. Given the small number of items remaining (after removing two raters), we decided to continue having all remaining passages double rated, and to compute reliability again using a greater number of items. Raters were not debriefed and did not meet together again. They continued rating the remaining passages, with each passage randomly distributed to two different raters, but with Raters 2 and 6 not having any passages in common. This was based on the preliminary reliability results, in anticipation of problems with Raters 2 and 6, and the potential need to completely drop all of their ratings. Table 5 displays the results of the reliability for all items.

Table 5. Reliability Results for the Cognitive Group for All Items

Feature	Overall Reliability	Removing Two Raters
Item type	0.70	0.82
Scope	0.70	0.83
Depth of knowledge	0.62	-

Note. For item type, Raters 3 and 6 were removed. For scope, Raters 2 and 3 were removed.

The overall reliability for each cognitive feature using all available items was again less than ideal. However, for Item Type, reliability increased to 0.82 after removing the ratings from Raters 3 and 6. The ratings for the five passages that were rated by only Raters 3 and 6 were then dropped and re-rated by Rater 4, who had the highest reliability for Item Type. For scope, reliability increased to 0.83 after removing the ratings from Raters 2 and 3. The ratings for the five passages that were rated by only Raters 2 and 3 were then dropped and re-rated by Rater 1, who had the highest reliability for scope.

For depth of knowledge, however, no removal of any one or two raters improved reliability. It was determined that raters would need to meet to achieve consensus and re-rate items. Since Raters 2 and 3 seemed most problematic, they were not invited back. Raters 4, 5, and 6 had the highest reliability as a group (0.72); however, Rater 5 was no longer available to work. The reliability of Raters 1, 4, and 6 as a group was 0.67. These three raters met and discussed the discrepancies among their shared passages (15 passages), and achieved verbal consensus. After achieving consensus, another random set of 18 passages were distributed evenly across the three raters, so that each rater received 12 passages and each passage was individually rated by two raters. Reliability was then computed, which was 0.715. A decision was made to have the 18 passages individually rated by all three raters, instead of just two. Reliability was computed and was 0.791 (reliability results are shown to the thousandths place here to demonstrate improvement). All remaining passages were subsequently rated for depth of knowledge by each of the three raters, including the 15 passages on which they had reached verbal consensus. Again, reliability was computed, which was 0.794, which was considered

acceptable. The ratings used in the final analyses were an average of "acceptable" ratings within each feature, when an item feature was rated by more than one "acceptable" rater.

For the Grammar Group, 18 passages were individually rated by two raters, similar to the process for the Cognitive Group, so that each individual rater received six passages. Table 6 presents reliability results for all six grammatical features.

Table 6. Reliability Results for the Grammar Group for Items from 18 Passages

Feature	Overall Reliability	Removing One Rater	Re-Rating
Passive verbs	0.85	-	-
Complex verbs	0.91	-	-
Relative clauses	0.69	0.83	-
Subordinate clauses	0.86	-	-
Noun phrases	0.81	-	0.87
Entities	0.88	-	-

Initial results indicated that rating reliability was less than ideal for relative clauses, and lower than expected for noun phrases. For relative clauses, removing Rater 5 increased reliability to 0.83. For noun phrases, a closer inspection revealed that Rater 4 was potentially under-counting, while Raters 2 and 6 were potentially over-counting. Therefore, all raters were given clarifications on the definition of noun phrases, and asked to re-examine their noun phrase counts. Reliability was re-computed, and reliability increased to 0.87 for noun phrases as shown in Table 6. A closer inspection of Rater 5 revealed under-counting of relative clauses, including mistaking relative clauses for subordinate clauses. The remaining passages were distributed evenly across all six raters for individual rating so that each rater received about nine additional passages to rate. Due to time and resource constraints, and the acceptable reliability results, the remaining passages were rated for grammatical complexity by only one rater each. However, the nine passages assigned to Rater 5 were also given to Rater 1 (4 passages) and Rater 2 (5 passages) to rate only relative clauses and subordinate clauses, and their ratings were entered in lieu of Rater 5's, even though Rater 5 did rate every feature. It was determined that having each rater rate all six grammatical features was helpful in recognizing each feature (e.g., one needs to recognize clauses in order to recognize entities as subjects), rather than assigning one grammatical feature per rater. For the initial 18 passages that had been rated by two raters each, Rater 5's ratings were deleted in favor of the other rater. All other grammatical features that were rated by two raters each were averaged for use in the final analyses.

The process of training and retraining raters with the decision to delete some ratings indicated the need to clearly specify rater requirements during recruitment followed by

sufficient training to optimize inter-rater reliability. A generalizability theory analysis of linguistic and cognitive ratings could further indicate ideal numbers of raters to use in subsequent studies of test accessibility for students with disabilities.

Results

Our analyses will focus on the research questions presented earlier:

1. How and to what extent does the cognitive complexity of reading assessments (item type, depth of knowledge, and scope) impact the performance of students with disabilities?

2. How and to what extent do textual/visual features of reading assessments (number of pages, words per page, typeface changes, point size changes, font changes, and unnecessary visuals) impact the performance of students with disabilities?

3. How and to what extent do lexical A features (number of words greater than 7 letters, number of relevant paragraphs, and number of words in items and relevant paragraphs) and lexical density B features (average lexical density and number of uncommon words) of reading assessments impact the performance of students with disabilities?

4. How and to what extent do grammatical features of reading assessments (subordinate clauses, complex verbs, passive voice, relative clauses, entities, and noun phrases) impact the performance of students with disabilities?

5. Among the five major categories of complexity features in the assessment, which category or categories most discriminate between students with disabilities and their peers (with no disabilities) in terms of their reading performance?

As indicated earlier, to answer these questions, the reading assessments from three states were rated on 21 accessibility features in five general categories: (1) cognitive complexity, (2) textual/visual complexity, (3) lexical A complexity, (4) lexical density B complexity, and (5) grammatical complexity. The cognitive complexity category included measures of passage and item types, depth of knowledge, and scope. The textual/visual complexity category included column count, number of pages, words per page, number of typeface changes, number of point size changes, number of font changes, and number of unnecessary visuals. The lexical A complexity category included a count of the number of words greater than seven letters in items and paragraphs, the number of relevant paragraphs, and the number of words in items and relevant paragraphs. The lexical density B complexity category included the average lexical density (total unique words per page/total words per page), and the number of uncommon words in items and relevant paragraphs. The grammatical complexity category included counts of the number of subordinate clauses, complex verbs, passive voice, relative clauses, entities, and noun phrases.

Two approaches were employed for analyzing the data: (1) a Multiple Discriminant (MD) approach, and (2) a Differential Item Functioning (DIF) approach. In the MD approach, we examined the differential impact of the accessibility features between SWDs and students without disabilities across the entire test; and in the DIF approach, Differential Bundle Functioning (DBF) and Differential Test Functioning (DTF) approaches were applied to see the impact of the accessibility features on the entire test as well as on the individual test items or a group of test items (bundle of items) that share specific accessibility features. Following is a report of the results of these analyses in two different sections, results of discriminant function analyses and results of DBF and DTF analyses.

Cognitive, Textual/Visual, Lexical A, Lexical Density B, and Grammatical Complexity Features of States' Reading Assessments

For each of the five categories, a summary of the descriptive statistics for each feature is presented in Tables 7 through 11. These summaries do not address student performance, but instead summarize characteristics of rated features. The data presented in these tables are averaged across the items in each assessment.

Cognitive. A summary of the cognitive ratings for all items is presented in Table 7. An informational item was coded with a score of zero, while an inferential item was coded with a score of one. In State A, items were more likely to be informational and less likely to be inferential than in the other states. About 40% of the State A items were considered inferential compared to over 60% of the items in the other assessments. The State A items also had a lower mean depth of knowledge score when compared to the other assessments. These results are interesting and somewhat counterintuitive when considering that the State A assessment tended to be more difficult than the other assessments for students with disabilities. The scope measure tended to be higher in State A when compared to the other two states.

Textual/Visual. Tables 8 and 9 present a summary of textual/visual characteristics of passages that are associated with the items that were analyzed. The data presented in each table is averaged across passages in each assessment. There was some variation in the passage format across the assessments. In State A each passage was presented in a single column format while State C generally had passages with text split into two columns. State B mixed single and two column text formats across the passages. The average number of pages of text tended to be somewhat lower for passages from State B when compared to the other two states. Text density as measured by the number of words per page was highest in State A with an average of over 450 words per page. As shown in Table 9, the number of unnecessary visuals and the number of typeface changes in passages also showed variation across the assessments. Items from State C were associated with passages that contained more unnecessarily visuals but fewer typeface changes than the other two states. Font-type and point-size changes were generally rare across the assessments.

Table 7. Descriptive Statistics of Cognitive Features at Item and Relevant Paragraph Levels

State	Assessment	Mean (SD)		
		Item Type	Depth of Knowledge	Scope
A	2006	0.40 (0.44)	1.40 (0.48)	2.27 (0.94)
A	2008	0.40 (0.46)	1.44 (0.44)	2.50 (1.04)
B	2006 F1	0.62 (0.46)	1.75 (0.51)	2.68 (1.18)
B	2006 F2	0.64 (0.41)	1.84 (0.55)	2.92 (1.19)
B	2006 F3	0.69 (0.42)	1.74 (0.44)	2.47 (0.88)
B	2006 F4	0.61 (0.43)	1.66 (0.48)	2.48 (0.96)
C	2006	0.66 (0.41)	1.75 (0.56)	2.31 (1.47)
C	2007	0.68 (0.41)	1.85 (0.54)	2.34 (1.44)
C	2008	0.68 (0.41)	1.75 (0.51)	2.22 (1.45)

Note. For State B, F1 through F4 refer to the four forms of the assessment.

Table 8. Descriptive Statistics of Textual/Visual Features at the Passage Level

State	Assessment	Mean (SD)		
		No. of Columns	No. of Pages	No. of Words Per Page
A	2006	1.00 (0.00)	1.73 (0.49)	473 (1.09)
A	2008	1.00 (0.00)	1.95 (0.32)	455 (1.27)
B	2006 F1	1.45 (0.50)	1.47 (1.00)	386 (172)
B	2006 F2	1.64 (0.48)	1.46 (0.50)	404 (117)
B	2006 F3	1.68 (0.47)	1.51 (0.52)	425 (168)
B	2006 F4	1.70 (0.46)	1.63 (0.78)	381 (120)
C	2006	1.95 (0.22)	1.73 (0.62)	418 (131)
C	2007	1.96 (0.19)	2.10 (0.81)	445 (113)
C	2008	1.86 (0.35)	1.55 (0.50)	426 (184)

Note. For State B, F1 through F4 refer to the four different forms of the assessment.

Table 9. Descriptive Statistics of Additional Textual/Visual Features at the Passage Level

State	Assessment	Mean (SD)				
		No. of Typeface Changes	No. of Font Changes	No. of Point-size Changes	No. of Visuals	No. of Unnecessary Visuals
A	2006	3.00 (2.11)	0.00 (0.00)	0.06 (0.17)	0.75 (0.44)	0.69 (0.43)
A	2008	4.75 (4.51)	0.13 (0.22)	0.31 (0.83)	0.75 (0.44)	0.69 (0.43)
B	2006 F1	3.95 (7.09)	0.00 (0.00)	0.56 (1.50)	1.20 (2.62)	0.00 (0.00)
B	2006 F2	2.82 (4.16)	0.00 (0.00)	0.13 (0.33)	0.95 (2.26)	0.09 (0.29)
B	2006 F3	5.86 (11.87)	0.38 (1.00)	0.31 (0.83)	1.13 (1.48)	0.63 (0.92)
B	2006 F4	2.97 (6.36)	0.00 (0.00)	0.13 (0.43)	1.34 (1.28)	0.79 (0.81)
C	2006	1.33 (1.84)	0.27 (0.42)	0.37 (0.40)	1.50 (0.94)	1.50 (0.94)
C	2007	2.78 (3.56)	0.06 (0.17)	0.13 (0.33)	1.00 (0.89)	0.98 (0.90)
C	2008	1.56 (1.93)	0.12 (0.21)	0.32 (0.33)	0.84 (0.87)	0.84 (0.87)

Note. For State B, F1 through F4 refer to the four different forms of the assessment.

Lexical. Table 10 presents a summary of the lexical ratings from each assessment. Lexical density B was calculated both for passages and item stems. For each item certain paragraphs were deemed relevant to obtaining the correct answer by raters. Means of the relevant number of paragraphs, the number of words, uncommon words, and words greater than seven are also presented in this table. There was substantial variation in the lexical characteristics across assessments. This variation occurred both within and between states. For example in the 2007 State C assessment an average item was associated with 3.86 relevant paragraphs compared to an average of 1.86 relevant paragraphs for the 2008 State C assessment. Items from the State B assessment in general were associated with a fewer number of words on average than the other states and also tended to have a lower proportion of those words exceeding seven letters. The degree of lexical density in passages was lowest in State A, while the degree of lexical density in item stems was lowest in State B.

Grammatical. Table 11 presents a summary of the grammatical ratings for items and their relevant paragraphs. There was a large degree of variation in the mean number of grammatical features identified across the assessments. This variation occurred both within and between states. The State C 2007 assessment had the highest average grammatical features count in items and their relevant paragraphs across all the features shown. The State C assessments in general also had more subordinate and relative clauses than the other assessments.

Table 10. Descriptive Statistics of Lexical A and Lexical Density B Features: Means and Standard Deviations (SD)

State	Assessment	Passage Level Lexical Density	Item Level Lexical Density	Item and Relevant Paragraph Level			
				No. of Relevant Paragraphs	No. of Words	No. of Uncommon Words	No. of Words >7 Letters
A	2006	0.45 (0.04)	0.83 (0.12)	2.28 (2.87)	115 (181)	50 (67)	23 (34)
A	2008	0.45 (0.05)	0.80 (0.13)	3.50 (5.59)	163 (250)	60 (73)	24 (27)
B	2006 F1	0.50 (0.09)	0.79 (0.13)	2.62 (3.76)	106 (126)	37 (32)	18 (15)
B	2006 F2	0.54 (0.09)	0.76 (0.12)	3.31 (4.67)	141 (131)	55 (42)	23 (16)
B	2006 F3	0.50 (0.08)	0.74 (0.13)	2.43 (5.34)	108 (125)	45 (37)	17 (13)
B	2006 F4	0.51 (0.08)	0.73 (0.12)	2.19 (3.16)	120 (156)	46 (38)	19 (16)
C	2006	0.51 (0.12)	0.82 (0.10)	2.59 (4.44)	187 (307)	64 (96)	28 (39)
C	2007	0.45 (0.12)	0.81 (0.10)	3.86 (6.39)	216 (334)	70 (90)	36 (45)
C	2008	0.53 (0.14)	0.81 (0.10)	1.86 (3.47)	153 (259)	48 (67)	22 (29)

Note. Lexical density aggregated across item and their relevant paragraphs was not available. For this reason we present this measure both for paragraphs and item stems. For State B, F1 through F4 refer to the four different forms of the assessment.

Table 11. Descriptive Statistics of Grammatical Features at Item and Relevant Paragraph Levels

State	Assessment	Mean (SD)				
		Passive Verbs	Complex Verb	Relative Clause	Subordinate Clause	No. of Entities
A	2006	1.41 (3.65)	1.54 (2.41)	2.31 (4.13)	2.65 (2.59)	10.0 (11.8)
A	2008	2.09 (3.91)	1.44 (2.28)	2.90 (4.40)	4.16 (5.05)	13.8 (17.1)
B	2006 F1	1.62 (2.27)	1.43 (2.07)	2.45 (2.32)	3.51 (4.03)	10.8 (10.4)
B	2006 F2	2.33 (4.21)	1.77 (1.92)	3.16 (2.87)	3.17 (2.65)	13.5 (10.5)
B	2006 F3	1.73 (1.68)	2.03 (3.18)	2.32 (2.31)	3.78 (3.76)	10.3 (10.3)
B	2006 F4	1.56 (1.39)	2.21 (3.02)	1.93 (2.21)	3.49 (4.99)	12.8 (13.1)
C	2006	1.85 (2.55)	1.49 (2.94)	3.89 (6.70)	4.86 (7.81)	15.9 (24.9)
C	2007	2.44 (3.41)	3.23 (5.60)	4.32 (6.01)	5.52 (8.87)	20.0 (26.8)
C	2008	1.59 (2.09)	1.64 (3.21)	2.76 (4.46)	4.75 (8.62)	11.9 (16.7)

Note. For State B, F1 through F4 refer to the four different forms of the assessment. Noun entities did not account for variation in results and the category was subsequently dropped from the analysis.

Differential Level of Impact of Accessibility Features on Reading Assessments for Students with Disabilities: Results from a Multiple Discriminant Analyses

There are many statistical models that can be used to examine the impact of the accessibility features on student performance. A multiple discriminant function provides a clear interpretation of such impact. This model provides a direct approach in comparing the performance of SWDs and students without disabilities in terms of the impact of the 21 accessibility features. Data from the three states were used for this study. A data file was created in which a student's incorrect response (0 score) in each test item was replaced with ratings from each of the corresponding features. Therefore, the total score of a particular feature for each student was the incorrect responses (0) plus the rating of the feature for each individual item. As a result, 21 scores for each student were created, with one for each accessibility feature. For example, using feature #2 (Item Type, as seen in Table 3), if a student responded to test item 1 incorrectly and item 1 had an item-type rating of 4, then the student's incorrect score on item 1 would be 4. A similar procedure was used for creating other feature scores, thus the units of analysis in this study were individual students not test items.

The next step of the analysis involved creating five composite scores from the 21 accessibility features. Table 3 presents a list of the 21 features along with their corresponding categories. In creating a composite score for each of the five categories, we first examined the assumption of unidimensionality for each of the five scales by conducting principal components (PC) analyses on each of the categories. After the unidimensionality assumption was confirmed, we then used the factor scores for each subscale to form the latent composite score for the subscale.

The analyses were conducted on data from two states, State A and State C. It was decided to use data from only one state to provide cross validation for the multiple discriminant analysis. To maintain consistency with State A's assessments across multiple years, State C was selected instead of using State B's 2006 test with four forms. We will discuss the results for each of these states separately, and the results will be compared and will serve as cross-validation data.

State A results. Tables 12 through 16 present the outcomes of the principal components analyses. As data in these tables show, the accessibility features within each category were highly internally consistent and all have high loadings with their corresponding factor (construct). For example, as data in Table 12 show, all three components for the cognitive complexity category (item type, depth of knowledge, and scope) had high loadings (.990 and above) with the overall cognitive complexity factor. The first component explained over 98% of the variance of the three individual features.

Table 12. Results of Principal Components Analyses for Cognitive Complexity Features: State A

	Component 1
Item type	.994
Depth of knowledge	.996
Scope	.990

Table 13 presents the results of principal components (PC) analyses for text features. Similar to what was presented for the cognitive complexity the six features under the "textual/visual complexity" category have near perfect factor loadings with the latent variable of this category (all factor loadings are at or above .950). The latent variable explained over 93% of the variance of the six features under this category.

Table 13. Results of Principal Components Analyses for Textual/Visual Features: State A

	Component 1
Number of pages	.976
Words per page	.988
Typeface changes	.962
Point-size changes	.976
Font-changes	.947
Unnecessary visuals	.950

Table 14 presents the results of principal components analyses for the lexical A category and Table 15 presents results for lexical density B features. Similar to the data presented above for other categories, all three accessibility features in the lexical A category and the two features in the lexical density B category have near perfect loadings with their respective latent variable. This latent variable explained over 90% of the common variance between the features.

Table 14. Results of Principal Components Analyses for Lexical A Features: State A

	Component 1
Number of words with more than seven letters	.984
Number of relevant paragraphs	.976
Number of words in item and relevant paragraphs	.920

Table 15. Results of Principal Components Analyses for the Lexical Density B Features: State A

	Component 1
Average lexical density	.971
Number of uncommon words in item and relevant paragraphs	.971

Finally, Table 16 presents the results of principal components analyses for the grammatical complexity category. Once again, the results of principal components analyses are quite similar with the three categories presented above. All six individual accessibility features in this analysis loaded nearly perfectly with the grammar latent variable, and this latent variable explains over 95% of the variance of the same six features.

Table 16. Results of Principal Components Analyses for Grammatical Features: State A

	Component 1
Subordinate clauses	.980
Complex verbs	.917
Passive voice	.977
Relative clauses	.994
Entities	.996
Noun phrases	.949

In this analysis, the presence of high levels of unidimensionality justified the creation of the five latent variables, one for each of the five categories of accessibility features. Factor scores for each of the five categories were computed and a multiple-discriminant analyses model was used in which the grouping variable was student's disability status (0 with no disability and 1 with disability). The multiple-discriminant model yielded one discriminant function with a Wilks' Lambda of .819, indicating that the function was highly significant. The canonical correlation for this model was 0.425 suggesting that the five latent variables explained over 18% of the variance in assessment outcomes according to students' disability status.

Table 17. Standardized Canonical Discriminant Function Coefficients: State A

	Function 1
Factor score for cognitive complexity	.531
Factor score for textual/visual complexity	1.657
Factor score for lexical A complexity	.781
Factor score for lexical density B complexity	.016
Factor score for grammatical complexity	.593

As the data in Table 17 suggest, the five latent accessibility variables each have quite a different impact on the performance of students according to their disability status. Among the five overall categories of features, the textual/visual features had the highest level of discrimination power between students with and without disabilities. Lexical A features constitute the next most powerful discriminating variable, while the other three categories (lexical density B complexity, grammatical complexity, and cognitive complexity) significantly contributed to the discriminant model yet do not show high power in discriminating between the two groups.

As indicated in Table 17, the textual/visual complexity component was the strongest discriminating variable between SWD and non-SWD. To further examine the contribution of this component on student performance, we removed this component from the discriminant model. Results of these analyses are reported in Table 18. As data in Table 18 indicate, removal of the textual/visual complexity component from the analyses showed other variables to have stronger power in discriminating between the two groups (SWD versus non-SWD). In other words, the strong impact of "textual/visual" complexity overwhelms the effects of other variables in predicting student performance. For example, the lexical density feature showed a small power coefficient in predicting differentiating between SWD and non-SWD when the textual/visual feature was included (.135); but the discriminant function coefficient for lexical density increased significantly (3.123) when textual features were removed from the model.

Table 18. Standardized Canonical Discriminant Function Coefficients when Textual/Visual Features are Excluded: State A

	Function 1
Cognitive complexity	-1.007
Lexical A	.032
Lexical density B	3.123
Grammatical	-1.195

State C results. Results of analyses of data from State C were very consistent with those from State A. Principal components analyses in all five accessibility categories yielded nearly perfect factor loadings indicating that the accessibility features within each category are highly internally consistent as indicated in Table 19.

Table 19. Results of Principal Components Analyses for Five Complexity Categories for State C

Complexity	Component 1
Cognitive features	
Item Type	.994
Depth of knowledge	.996
Scope	.990
Textual/visual features	
Number of pages	.976
Words per page	.988
Typeface changes	.962
Point-size changes	.976
Font changes	.947
Unnecessary visuals	.950
Lexical A features	
No. of words with more than seven letters	.984
No. of relevant paragraphs	.976
No. of words in item and relevant paragraphs	.920
Lexical density B features	
Average lexical density	.971
No. of uncommon words in item and rel paragraphs	.971
Grammar	
Subordinate clauses	.980
Complex verbs	.917
Passive voice	.977
Relative clauses	.994
Entities	.996
Noun phrases	.949

Results from analyses of data from State C were used to cross-validate findings from State A. As data in Tables 12-16 for State A and data in Table 19 for State C show, the factor loading on the five accessibility constructs are quite similar and in most cases almost identical. Similarly, as data in Tables 17-18 for State A and Tables 20-21 for State C show, the pattern of results from the discriminant analyses are also quite consistent. These consistencies between the outcomes of analyses from two independent states provide cross-validation evidence for this study.

The outcomes of discriminant analyses for State C are shown in Table 20.

Table 20. Standardized Canonical Discriminant Function Coefficients: State C

	Function 1
Factor score for cognitive complexity	.145
Factor score for textual/visual complexity	1.147
Factor score for lexical A complexity	.198
Factor score for lexical density B complexity	.135
Factor score for grammatical complexity	.043

As data in Table 20 show, the results of discriminant analyses for State C are reasonably consistent with the results for State A. Among the five categories, textual/visual complexity has the highest level of impact on the performance of students with disabilities and this category has the highest power in discriminating between these students and their peers with no disabilities.

Table 21. Standardized Canonical Discriminant Function. Coefficients when Textual/ Visual Features are Excluded: State C

	Function 1
Cognitive complexity	.114
Lexical A complexity	-.948
Lexical density B	2.495
Grammatical	-.740

Similar to what was reported for State A, when textual/visual features were excluded as shown in Table 21, then other features showed discrimination power. For example, Lexical Density B shows substantial discrimination power (2.495) in the absence of textual/visual features.

A major limitation of this study is that the results of the discriminant analyses of the accessibility features may be confounded with students' disability status. That is, a canonical correlation of .425 may explain the overall performance difference between students with disabilities and other students; however, the differential level of impact of the five latent accessibility features suggests that some of these features can clearly have more impact on the performance of students with disabilities than other features.

Differential Level of Impact of Accessibility Features on Reading Assessments for Students With Disabilities: Results from Differential Item Functioning Using the Non-Compensatory Differential Item Functioning (NCDIF) Index and Logistic Regression

Results from discriminant analyses indicated that some of the accessibility features have more impact on student outcomes particularly for students with disabilities. These results can be interpreted at the total test level. However, we also wanted to know whether some of the test items are more affected by some of these features than other items. We therefore conducted a series of Differential Item Functioning (DIF) and Differential Test Functioning (DTF) analyses. Results of these analyses are presented in this section of the report.

DIF is said to be present for an item when the probability of answering an item correctly is different between two groups who have the same performance on the total test. A number of techniques to detect DIF have been proposed including Mantel-Haenszel (Holland & Thayer, 1988), standardization procedure (Dorans & Kulick, 1983, 1986), SIBTEST (Shealy & Stout, 1993), and logistic regression approach (Spray & Carlson, 1988). Through a simulation study, Leon (2009) found that when focal group ability and population size differed substantially from the reference group such as SWDs and non-SWDs groups, a DIF and DBF detection method that combined logistic regression with the NCDIF index showed superior power with fewer Type I errors when compared to other methods such as Mantel Haenzsel, SIBTEST and crossing SIBTEST. The conditions tested in the simulation study were intended to mimic those seen in populations of ELLs and students with disabilities.

The DIF and DBF techniques we employ in this paper make use of the NCDIF index combined with logistic regression and reflect the recommendations of the previously referenced simulation study (Leon, 2009). Logistic regression is used to obtain the predicted probability of a correct answer for each student based on his or her group membership and score on a matching ability criterion. The Non-Compensatory Differential Item Functioning index (NCDIF) (Raju, van der Linden, & Fleer, 1995) is a

method for detecting DIF and measuring its magnitude. NCDIF is used to measure the area between the two item-characteristic curves for the focal and reference groups that were obtained with logistic regression. The matching criterion for the focal and reference groups is not simply the total assessment score. Instead, items with larger DIF magnitude are given less weight in the matching criterion than items with smaller DIF magnitude. This process helps to remove DIF contamination from the matching criterion. The NCDIF formula can be expressed as follows:

NCDIF= $\sigma_d^2 + \bar{\mu}_d^2$

Where *d* represents the difference between the expected score for a focal group subject from what the expected score would have been for a reference group subject of the same ability.

As shown in the formula, the NCDIF index can be separated into two components. The mean difference between the focal and reference groups squared ($\bar{\mu}_d^2$) can be considered a measure of the magnitude of uniform DIF. Similarly, the variance of the estimated probability differences (σ_d^2) can be considered a measure of the magnitude of non-uniform DIF. In addition, the mean difference between the focal and reference groups ($\bar{\mu}_d$) can serve as a signed measure of uniform DIF.

With respect to cut-points to define DIF magnitudes as moderate and large DIF, a standard cutoff value for NCDIF of .006 has been proposed for use with binary items (Teresi et al., 2007). Under the SIBTEST technique, moderate DIF magnitude is reached when the absolute value of the beta-uni statistic is >0.059 and large DIF magnitude is reached when the absolute value of the beta-uni statistic is >0.089. In a simulation study (Leon, 2009), comparable cut-points for the NCDIF index used with logistic regression were found at >.0082 for moderate DIF, and 0.0136 for large DIF. In this study, we define DIF at the moderate cut-point of the NCDIF index with logistic regression (>0.0082).

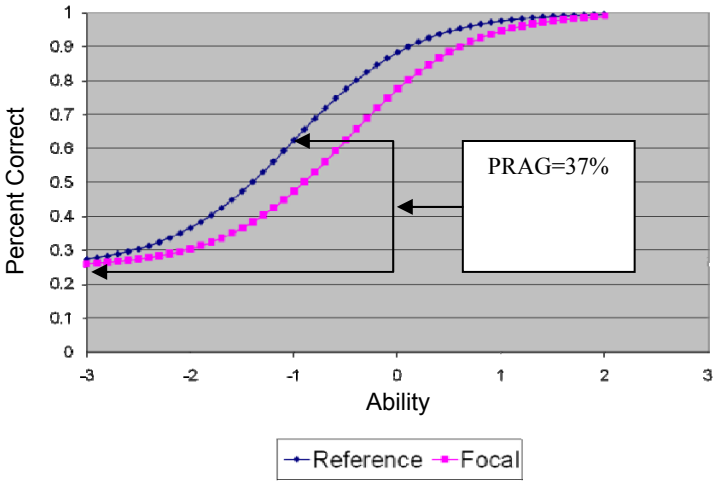
Differential test and differential bundle functioning. Raju has also shown that differential test functioning (DTF) can be modeled just as DIF was at the item level with the NCDIF index. Rather than focusing on individual item functioning, DTF is a measure of how well an assessment functions across all of its items. Similarly, differential bundle functioning (DBF) can be modeled like DTF but for groups of items (or, item bundles) rather than for the entire test (McCarty, Oshima, & Raju, 2007). The differential functioning method employed in this study (i.e., NCDIF index combined with logistic regression) allows not only for analysis at the item level (i.e., DIF) but also at the item bundle level (i.e., DBF). The advantage of DBF analyses is that items within an assessment can be grouped based on a theoretical framework, and that framework can be analyzed for both uniform and non-uniform functioning. As discussed earlier in this report, data for this study were obtained from three states. The data included test items as well as students' performances on the items.

Item difficulty impact on signed uniform DIF ($\bar{\mu}_d$). In this study we examined

how potential nuisance factors (such as poor readability, unnecessary visuals, etc.) may impact test performance for students with disabilities. Signed uniform DIF ($\bar{\mu}_d$) is a measure specific to a population group of interest, in this case SWDs. Certain psychometric characteristics of test items can greatly influence how a potential nuisance factor impacts signed uniform DIF for groups of students who have large differences in ability. To illustrate this, Figures 1 and 2 show theoretical examples of an "easy" item and a "difficult" item, respectively. The item characteristic curves for two groups of hypothetical students are shown – a lower ability focal group and an average ability reference group. These groups were chosen to illustrate a possible group of students with disabilities (focal group) and students without disabilities (reference group).

In Figure 1, the IRT-based discrimination parameter was set at 1.0 for both groups and the guessing parameter was set at 0.25 (assuming that there are only four choices in the item). The location (DIF) parameter was set to -1.0 for the reference group and to -0.5 for the focal group (to indicate an easy item). Based on these parameters, a reference group student with the same ability as an average focal group student (-1 *SD*, or one standard deviation below the mean) would be predicted to get this item correct about 62% of the time, while the average focal group student would get the item correct about 48% of the time. The 0.5 difference for this item in the location parameter results in a signed uniform DIF score ($\bar{\mu}_d \bar{\mu}_d$) of about -0.10. For the item in Figure 1 the percentage range above guessing (PRAG) for a reference group student at the "focal group mean" is 37 percentage points. This was computed by subtracting 25% (the percentage students would get correct by simply guessing among four choices) from 62%. (62-25=37).

Figure 1. Theoretical example of an item characteristic curve of an "easy" item

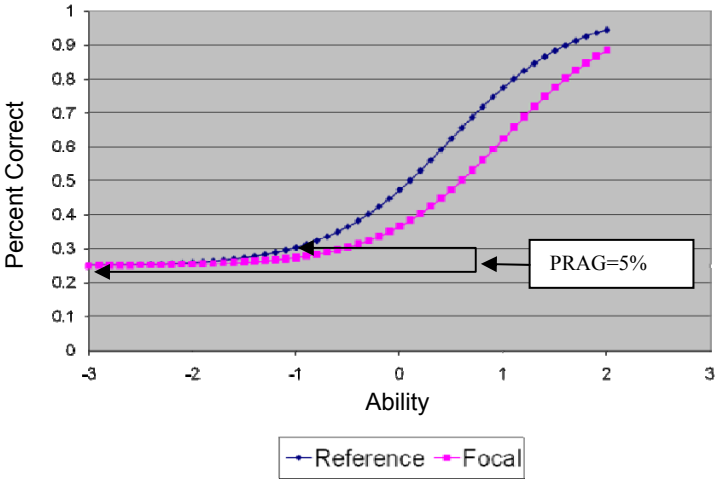


Note: If the guessing parameter is set at 25%, the percentage range above guessing (PRAG) is 37 percentage points for a reference group student whose "ability" is 1 standard deviation below the mean.

Figure 2, as a comparison, shows an item characteristic curve comparison for a "difficult" item. The IRT-based discrimination parameter was set at 1.0 for both groups, and the

guessing parameter was set at 0.25, the same as for the item example shown in Figure 1. In contrast, the location (DIF) parameter was set to 0.5 for the reference group and to 1 for the focal group (to indicate a difficult item). A reference group student with the same ability as an average focal group student (-1 *SD*, or one standard deviation below the mean) would now be predicted to get this item correct about 30% of the time, while the average focal group student would get the item correct about 27% of the time. The same 0.5 difference in the location parameter set for the easy item illustrated earlier, results in a signed uniform DIF score ($\bar{\mu}_d \bar{\mu}_d$) of less than -0.05 for the difficult item in this example. This occurs because in the Figure 2 example, there are relatively few focal group students represented over the ability range where the DIF occurs. For the item shown in Figure 1, the percentage range above guessing (PRAG) for a reference student at the focal group mean was much higher (62-25=37) as compared to the item shown in Figure 2 (30-25=5). The practical impact on student performance of the nuisance factor (i.e. cognitive, lexical A, lexical B density, grammatical, or textual/visual complexity features) is much less for the item in Figure 2 than the item in Figure 1. In other words, a nuisance factor will have less impact on student performance when an item is very difficult.

Figure 2. Theoretical example of item characteristic curve of a "difficult" item



Note: If the guessing parameter is set at 25%, the percentage range above guessing (PRAG) is 5 percentage points for a reference group student whose "ability" is 1 standard deviation below the mean.

The degree to which an item discriminates can also impact the practical influence of a potential nuisance factor. For instance, if the example item shown in Figure 1 had a discrimination parameter equal to 0.5 instead of 1.0, the signed uniform DIF score ($\bar{\mu}_d \bar{\mu}_d$) would be reduced from about -0.10 to -0.07. The item PRAG however, appears to be the limiting factor. If the example item shown in Figure 2 had a discrimination parameter equal to 0.5 instead of 1.0, the signed uniform DIF score ($\bar{\mu}_d \bar{\mu}_d$) would remain the same (just below -0.05). The presence of a nuisance factor will be difficult to detect in an item that has a small PRAG due to content difficulty, and the focal group of interest would score low on the matching ability criterion. The presence of a

nuisance factor in an item that has a large PRAG and discriminates well within a low-ability population, however, will likely have a greater impact on the total score and be detected with DIF analyses. Knowing that characteristics such as item PRAG and item discrimination can influence the signed uniform DIF measure, we are careful to control for these item characteristics in our analyses in order to isolate the nuisance variable effects.

Results of analyses of differential item and differential test functioning. Results for both DIF and DTF are summarized in Tables 22 and 23 for the SWD and SLD groups respectively. A negative sign associated with the signed uniform DTF measure ($\bar{\mu}_D\bar{\mu}_D$) indicates that the test favors the reference group (i.e., the test functions against the focal group). For example, in the State B 2006 Form 1 assessment, the $\bar{\mu}_D\bar{\mu}_D$= -1.10 indicates that an average SWD is predicted to score 1.10 raw score points lower than a non-SWD student of similar reading ability on the matching criterion. The larger magnitude of signed uniform DTF against the focal group tended to be loosely associated with assessments that had high PRAG averages. The State A assessments had a lower PRAG indicating that these assessments were difficult for an average non-SWD student equal in ability to the median SWD student. There were fewer DIF items identified and the magnitude of signed uniform DTF against the focal group was also smaller in general on the State A assessment. The average PRAG was higher in States B and C than in State A. Nevertheless there was significant variation in the number of DIF items identified in States B and C with more items generally identified as DIF in State C. We will later explore whether differences in DIF findings not due to PRAG or discrimination can be explained in part by differences in various features.

Interestingly, State A had the highest point-biserial correlations despite having the lowest PRAG. This may seem counterintuitive, but as shown in Table 22 despite fewer test items and despite answering fewer items correctly State A SWD students had a standard deviation as large as the other states. The SLD analyses presented in Table 23 only include the seven assessments from States B and C because information on disability type was not available in State A. In general the SLD results for States B and C were similar to the SWD results.

Table 22. Differential Item and Differential Test Functioning Across State Assessments-A Comparison of SWD to non-SWD

State	Reading Assessment	No. of Items	Average PRAG	Average Point-Biserial Discrimination	No. of Items Detected as DIF	Signed Uniform DTF
A	2006	48	11	0.41	5 (10%)	-0.64
A	2008	48	13	0.41	1 (2%)	-0.02
B	2006 F1	56	25	0.34	8 (14%)	-1.10
B	2006 F2	56	26	0.34	7 (13%)	-0.93
B	2006 F3	56	28	0.32	6 (11%)	-1.07
B	2006 F4	56	25	0.33	11 (20%)	-1.09
C	2006	58	23	0.33	17 (29%)	-1.38
C	2007	56	21	0.34	20 (36%)	-1.34
C	2008	56	22	0.35	17 (30%)	-1.54

Note. Percentage indicates the percentage of DIF items out of the total number of items per test. For State B, F1 through F4 refer to the four different forms of the assessment.

Table 23. Differential Item and Differential Test Functioning across States Assessments-A Comparison of SWD to Students with SLD

State	Reading Assessment	No. of Items	Average PRAG	Average Point-Biserial Discrimination	No. of Items Detected as DIF	Signed Uniform DTF
B	2006 F1	56	25	0.34	6 (11%)	-1.06
B	2006 F2	56	26	0.34	7 (13%)	-0.70
B	2006 F3	56	28	0.32	6 (11%)	-0.94
B	2006 F4	56	25	0.33	8 (14%)	-0.79
C	2006	58	23	0.33	20 (34%)	-1.54
C	2007	56	21	0.34	21 (38%)	-1.11
C	2008	56	22	0.35	20 (36%)	-1.56

Note. Percentage indicates the percentage of DIF items out of the total number of items per test. For State B, F1 through F4 refer to the four different forms of the assessment.

Regression models. A multiple regression approach was applied in order to examine the relationship between each of the complexity features and the signed uniform DIF ($\bar{\mu}_d$) findings. In our first set of analyses we examined the relationships of each individual complexity feature to signed uniform DIF findings (see Table 24). Next we constructed a more comprehensive model that included measures from each of our complexity categories (see Table 3). These analyses were conducted at the item level across all nine reading assessments. As anticipated there was a strong correlation between item PRAG and the signed uniform DIF results (r = -0.762). The majority of items indicating

DIF against SWDs had PRAG values over 30. The data were split into three strata representing items with low PRAG (0-11), moderate PRAG (12-29) and high PRAG (30 or above). In each model the continuous measure of signed uniform DIF served as the dependent variable. A two-step approach was used within each stratum. In step one, two psychometric measures, PRAG and the bi-serial correlation across the ability range of the focal group (Discrim) were entered as controls. In step two, measures from each of our complexity categories were entered. The change in R-square from step 1 to step 2 was examined to determine the amount of variance explained by the features captured by our rubric after controlling for the psychometric characteristics.

Individual complexity features. Table 24 shows the resulting change in R-square from step 1 to step 2 for each complexity feature within the five categories (cognitive, grammatical, lexical A, lexical density B, and textual/visual). Individual complexity features were much more likely to significantly contribute to the explanation of the variation in the DIF findings for the high PRAG items when compared to the low and moderate PRAG items. Of the 21 features modeled, 15 made significant contributions in the high PRAG items while only one feature had a significant r-square change within both the low and moderate PRAG items. Each of the 15 significant features in the high PRAG items had model coefficients in the expected direction. The strongest individual cognitive feature was depth of knowledge. Among the grammar features, complex verbs and subordinate clauses made the largest contributions. The more complex verbs and subordinate clauses that were present in items and their relevant paragraphs the more DIF was likely to be found against students with disabilities. Lexical density at the passage level and words greater in length than seven letters that were present in items and their relevant paragraphs were each also strongly related to the DIF findings. Finally a number of passage level textual/visual features were also significantly related to the DIF findings. Among those the strongest features were point size and font changes along with the number of unnecessary visuals. Additionally, it is interesting that the scope feature, which had a strong correlation to both the depth of knowledge and grammar features, was not significant when considered individually. We will examine next whether some of these features might help to explain the DIF findings in a multivariate context.

Table 24. Contribution of Individual Features to Explanation of DIF

Section and Feature	Low PRAG Change in R-square	Mod PRAG Change in R-square	Hi PRAG Change in R-square
Cognitive Features – Item level			
Item Type	0.012	0.000	0.018*
Depth of Knowledge	0.012	0.006	0.033**
Scope	0.000	0.002	0.001
Grammatical Features – Item and relevant paragraphs			
Passive Voice	0.003	0.009	0.007
Complex Verbs	0.002	0.012	0.052**
Relative Clauses	0.000	0.001	0.026*
Subordinate Clauses	0.003	0.000	0.046**
Noun Phrases	0.000	0.005	0.008
Entities	0.000	0.001	0.027*
Lexical B Density Features– Passage and item level			
Lexical density-passage	0.000	0.010	0.050**
Lexical density-item	0.005	0.005	0.017
Lexical A Features - Item and relevant paragraphs			
Total number of words	0.001	0.001	0.025*
Total number of words with 7+ letters	0.000	0.003	0.035**
Total number of uncommon words	0.002	0.003	0.018*
Textual/Visual – Passage level			
Columns	0.007	0.000	0.021*
Number of pages	0.003	0.011	0.023*
Words per page	0.001	0.005	0.000
Typeface changes	0.009	0.000	0.002
Font changes	0.004	0.009	0.053**
Point size changes	0.000	0.009	0.061**
Unnecessary visuals	0.022*	0.000	0.039**

Note. One asterisk is significant at .05 and two asterisks are significant at .01.

Mutivariate results (multiple complexity features). We use a multivariate approach to examine whether unique contributions to DIF were present across the five complexity categories and multiple features. To reduce the number of predictor variables in the model when multiple strong individual predictors are present within a feature, we combine those predictors into a latent feature factor. For example, from the grammatical section, complex verbs and subordinate clauses were combined into a single latent grammar measure (GRAMMAR) comprised of their shared variation. Similarly, a latent

measure (LEXICAL) combining lexical density at the passage level and words greater in length than seven letters that were present in items and their relevant paragraphs was used to represent the lexical section. The textual/visual section was represented by a latent variable (TEXTVIS) comprised of unnecessary visuals, point size changes, and font changes. Only depth of knowledge was a strong individual predictor from the cognitive section. The scope measure was not a strong individual predictor but appeared to pick up on cognitive aspects outside of depth of knowledge on investigation of correlations. In addition we include lexical density (lexical_item) at the item stem level to see whether an item stem specific measure might also contribute to the understanding of DIF results. A summary of the eight latent variables and components of which they are comprised is shown in Table 25.

Table 25. Differential Item and Differential Test Functioning Latent Variables

Latent Variables	Components
PRAG	percent range above guessing
DISCRIM	item discrimination
GRAMMAR	complex verbs, subordinate clauses
LEX (LEXICAL)	lexical density at the passage level words > 7 letters in items and relevant paragraphs
TEXTVIS	unnecessary visuals, point size and font changes
Depth of knowledge	individual predictor
Lexical_item	lexical density at the item stem level
Scope	cognitive component

After selecting the variables and prior to running the multivariate regression models, a principal components analysis with varimax rotation was conducted to examine the degree of multicollinearity present in the eight independent variables. The first principle component explained 26.9 percent of the variance. The largest single correlation between any two variables was r =.575 between TEXTVIS and GRAMMAR followed by r =.542 between scope and GRAMMAR.

The results from the regression analysis between the five complexity categories and their features and signed uniform DIF ($\bar{\mu}_d$) values are presented in Table 26. The change in R-square due to the addition of the features increases across the 3 strata as PRAG increases. For the Low PRAG items, the addition of the rubric features results in an R-square change of 0.024. The R-square change increases slightly for the Moderate PRAG items (R-square change = 0.040) and substantially for the High PRAG items (R-square change=0.232). This increased R-square change in the highest PRAG items is consistent with the expectation that these items would be more sensitive to detect the presence of potential DIF. Conversion of the High PRAG to Cohen's *d* yields *d*=1.1 which exceeds the general rule of thumb for a large effect size.

Table 26. Change in R-Square Due to Latent and Single Factors, Controlling for PRAG and Item Discrimination

Model	Variables Included	R	R-square	R- square Change	F Change	Sig. F Change
1 Low PRAG	PRAG & Discrim	.561	.315	.315	40.168	.000
	Plus Latent and Single Factors	.582	.338	.024	1.013	.419
2 Mod PRAG	PRAG & Discrim	.489	.239	.239	22.516	.000
	Plus Latent and Single Factors	.529	.280	.040	1.275	.273
3 High PRAG	PRAG & Discrim	.464	.216	.216	21.987	.000
	Plus Latent and Single Factors	.617	.448	.232	10.784	.000

In Table 27 the full model coefficients are presented from the High PRAG strata. A negative value in the dependent variable (signed uniform DIF) indicates DIF in the direction against SWD. Seven of the eight variables are significantly related to the dependent variable. Among the six complexity variables examined, five have negative coefficients indicating increased DIF against SWD with higher values of the features. Therefore as the values of GRAMMAR, depth of knowledge, TEXTVIS, and lexical density (LEXICAL and lexical_item) increase, an item in the High PRAG category is more likely to exhibit DIF against SWDs. The scope variable has a positive coefficient; a result that is not consistent with the rest of our findings. The TEXTVIS and scope measures were the two features that made the largest contribution toward explaining the variation in the DIF outcome.

Table 27. Final Regression Model Results from the High PRAG Strata (30+)

	Model	Unstandardized Coefficients		Standardized	t	Sig.
		B	*SE*	Beta		
Percent Above Guessing (30+)	(Constant)	.135	.034		4.007	.000
	PRAG	-.167	.026	-.429	-6.536	.000
	DISCRIM	-.127	.040	-.233	-3.202	.002
	Scope	.009	.003	.277	3.456	.001
	DepthKnowl	-.019	.005	-.247	-3.497	.001
	LEXICAL	-.004	.004	-.105	-1.149	.252
	GRAMMAR	-.006	.003	-.198	-42.053	.042
	TEXTVIS	-.009	.002	-.272	-4.335	.000
	Lexical_item	-.038	.019	-.121	-1.954	.052

Note. The dependent variable is signed uniform DIF. PRAG refers to the percentage range above guessing.

Discussion

Legislation mandates the inclusion of students with disabilities in large-scale state assessments, and thus most of these children participate in statewide assessments with or without accommodations. However, inclusion of students with disabilities in large-scale state assessments is based on the assumption that state assessments are appropriate and that the outcomes of these assessments are reliable, valid and fair for these students. In fact, there is some evidence that students with disabilities may do poorly on state and national assessments because the assessment does not address their disabilities. Assessments that have been developed for and normed on mainstream students may fail to present an accurate portrayal of knowledge and skills of students with disabilities.

Based on the literature cited in this report, students with disabilities perform substantially lower on standardized tests than students with no apparent disabilities in both state (Abedi, Leon, & Mirocha, 2003; Altman et al., 2009; Ysseldyke et al., 1998) and national assessments (Lee, Grigg, & Donahue, 2007). While part of this low performance may be explained by a student's specific disabilities or lack of access to the general education curriculum, a major part of it may be attributed to the limitations of existing state assessments in addressing the needs of these students. That is, a substantial part of the performance difference between students with disabilities and their peers without disabilities may be explained by accessibility issues. Current state assessments may not be sensitive enough to the needs and backgrounds of students with disabilities.

Based on the review of existing literature and consultations with experts in the field, we identified 21 accessibility features that could have major impact on the assessment outcomes of students with disabilities. These 21 features were grouped into the following five categories: (1) cognitive, (2) lexical A, (3) lexical density B, (4) textual/visual, and (5) grammatical. The grouping of the 21 features into five categories seems to be conceptually and analytically sound. Experts confirmed these categorizations and results of factor analyses of the features within each category yielded strong evidence of internal consistency of the features within the five categories.

Two different analytical approaches were used in this study; a differential item functioning (DIF) approach and a discriminant analysis approach. In the DIF approach (using DTF and DBF methods) sets of test items representing a particular accessibility feature were compared across groups formed by students' disability status. Groups of accessibility features that behaved differentially across the two groups were identified, and the level of impact on student reading performance was examined by their disability status in a multiple regression model. This multiple regression approach that examined the relationship between each complexity feature and the signed uniform DIF ($\bar{\mu}_d \bar{\mu}_d$) indicated that as item difficulty increased, 15 features were significantly associated with lower outcomes for students with disabilities than students without disabilities. These features include:

- Cognitive complexity—depth of knowledge.
- Grammatical complexity—complex verbs and subordinate clauses
- Lexical density B complexity—percent of unique words compared to total words at the passage level
- Lexical A complexity—words greater in length than seven letters present in items and their relevant paragraphs
- Textual/visual features—point size and font change, and number of unnecessary visuals.

In the discriminant analyses model, the latent scores of the five overall accessibility features were used as discriminating variables to identify the features that mostly discriminate students with disabilities from students without disabilities. This analysis suggests that textual/visual features have the highest level of discrimination power between students with disabilities and students without disabilities. The lexical A category was the second most powerful discriminating variable.

Results of these two analyses consistently suggested that: (1) some of the accessibility features had more impact on reading than other features, and (2) some of these features had more differentiating powers between students with disabilities than other features.

DIF Results

The results of DIF analyses showed that the pattern of DIF (testwide DIF, reported as DTF) varied across assessments and states. Data from different states showed different patterns of test functioning. Even within the states, different reading assessments showed different DIF patterns. These findings were expected because different states measure diverse content using various test items with distinct formats. The results also showed substantial variation in the complexity feature ratings across the assessments and states. Once again, this is quite expected as state assessments' goals and tools are unique.

Items with adequate percentage ranges above guessing (PRAG) showed features that are significantly associated with the DIF results, indicating that the performance of students with disabilities is more affected by these particular features. The results indicated that 15 out of the 21 accessibility features showed significant contributions among the high PRAG items. In fact, each of the 15 significant features in the high PRAG items had model coefficients indicating students with disabilities were more affected by these features.

The results also clearly indicated that each of the features offered unique contributions in explaining the DIF results. However, the associations between the five complexity categories and the DIF findings were not present in the items with moderate and low PRAG. It appears that the difficulty of these items impairs the detection of nuisance factors through signed uniform DIF.

It is important to note, however, that the results of DIF and discriminant analyses clearly show overlaps between features. Not only are the individual features highly correlated, but there is also some redundancy in the information provided by the five general categories of accessibility features.

Discriminant Results

Consistent with the results of the DIF analyses summarized above, the results of discriminant analyses also indicated that (1) some of the accessibility features had more impact on reading than other features, and (2) some of these features had stronger differentiating powers between students with disabilities and their peers without disabilities. A distinctive feature of the discriminant analyses outcomes is their consistency over the cross validation samples. As discussed in the results section, we used data from different states to serve as cross-validation data for each other. Even though there were some differences across states on the content and format of test items, the overall content and goals of measurement remained the same across the participating states in this study. More importantly, consistencies between states on the impact of features, in spite of their differences, provided supporting evidence on the validity of results.

The results of discriminate analyses indicated that while all categories of features had impact on student performance, some categories showed a much greater level of impact. More importantly, some of these categories had more power in identifying which features had the highest level of impact on performance of students with disabilities.

Identifying features with the highest level of impact on the performance of students with disabilities has major implications for the assessment of SWD, particularly when the features could be easily altered without changing the construct measured. There are many factors that affect student performance on assessments, and some of these are essential components of the measures, such as the content and construct being measured. These cannot be altered because such changes might alter the construct being measured. However, some of these factors are incidental to the assessment and can be altered without having major impact on the outcome of measurement. For example, students with disabilities may find crowded test pages difficult and may experience fatigue and frustration when answering these items. Changing the test to include better readability for students does nothing to alter the construct, yet may significantly increase the performance of students with disabilities on such assessment items.

Conclusions

Based on expert judgment, the five overall categories of accessibility features were grouped into two major categories; Group A, those that are incidental to the measurement and can be modified, and Group B, those that are essential to the construct being measured and cannot be modified. Group A includes textual/visual and lexical A categories (see Table 3). Group B includes cognitive, lexical density B, and grammatical

complexity categories. As noted earlier, the determination to use two lexical categories was made based on expert opinion and research group consensus that lexical A features may have less serious impact than lexical density B features on reading constructs if the complexity is reduced. Because the target of measurement is reading, any of the Group B features may have more serious impact on the construct than Group A features and may alter it. These Group B features could possibly be revised without changing constructs, but there currently is no system in place to do so. Future research might explore how to better address Groups A and B alterations without changing the construct. For example, a future empirical study can include a field test with reduced lexical A and lexical density B complexity features presented to non-SWD groups. Construct validity decisions concerning these features may then be clearer based on performance outcomes of control and experimental non-SWD groups.

The results of discriminant analyses, which were consistent across the sites and test booklets, suggest that textual/visual as well as lexical A features have more impact on the performance of students with disabilities than the other features measured. Modifying these features on assessments may make them more accessible to students with disabilities without affecting the construct being measured.

Our findings suggest that there are certain revisions that can be done on the current assessments to make them more accessible for students with disabilities. These changes, particularly those under Group A features may be done relatively easily without altering the reading construct. Features such as words per page, typeface changes, point size changes, and unnecessary visuals may be easily adjusted to the optimum level without affecting the validity of the assessment. These changes may help all test takers, particularly students with disabilities who may become frustrated with excessive use of these features.

Findings of this study have informed us that even minor revisions of test content and format that do not alter the construct could be of great value in making assessments more accessible to all students, particularly those with disabilities. We plan to conduct a follow-up study (or series of studies) to examine the impact of each of these accessibility features by assigning students (both those with disabilities and those without apparent disabilities) to different levels of the features in a computer-based assessment system. Versions of assessments with different levels of visual complexities could be randomly assigned to students and the impact of this feature on the assessment outcomes can be determined.

The main concern in including the five accessibility features examined in this study in assessment systems for students with disabilities is the possibility that such features may alter the construct. This is a validity concern which can be addressed by assigning students without disabilities randomly to treatment and control conditions in which these students receive some or none of these features. Any change in their performance due to the possible impact of these features can then be determined.

In summary, the results of this study can help the assessment community in two ways. First, by elaborating on some test accessibility features, this report may serve as a guideline for those who are involved in test development and the instruction and assessment of students with disabilities. Second, and more importantly, this report provides methodology for examining other features that may have a major impact on assessment outcomes for students with disabilities.

References

Abedi, A., Leon, S., & Kao, J. (2008a). *Examining differential distractor functioning in reading assessments for students with disabilities* (CRESST Rep. No. 743). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, A., Leon, S., & Kao, J. (2008b). *Examining differential item functioning in reading assessments for students with disabilities* (CRESST Rep. No. 744). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abrahamsen, E. P., & Shelton, K. C. (1989). Reading comprehension in adolescents with learning disabilities: Semantic and syntactic effects. *Journal of Learning Disabilities, 22,* 569-572.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Altman, J., Thurlow M., & Vang. V. (2009). *Annual performance report: 2006-2007 state assessment data.* Minneapolis MN: National Center on Educational Outcomes, University of Minnesota. Retrieved from http://www.cehd.umn.edu/NCEO/onlinepubs/CharacteristicsBrief/default.htm

Alvermann, D. E. (1981, December). *Reading achievement and linguistic stages: A comparison of disabled readers and Chomsky's 6- to 10-year-olds*. Paper presented at the annual meeting of the National Reading Conference, Dallas, TX.

Artiles, A. (2003). Special education's changing identity: paradoxes and dilemmas in views of culture and space. *Harvard Educational Review, 73*(2), 165-202.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, A. L., Butler, F. A., & Sato, E. (2005). *Exploring common language demands in ELD and science standards.* (CSE Tech. Rep. No. 667). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bauman, J., & Culligan, B. (1995). *About the general service list*. Retrieved from http://jbauman.com/aboutgsl.html

Baxter, G. P., & Glaser, R. (1997). *An approach to analyzing the cognitive complexity of science performance assessments* (CSE Tech. Rep. No. 452). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Bloodsworth, J. G. (1993). *Legibility of Print*. Retrieved from ERIC database. (ED355497).

Botel, M. M. (1972). A formula for measuring syntactic complexity: A directional effort. *Elementary English, 49,* 513-516.

Butler, F. A., Bailey, A. L., Stevens, R. A., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks* (CSE Tech. Rep. No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Cain, K. (1996). Story knowledge and comprehension skill. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 167-192). Hillsdale, NJ: Erlbaum.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing: An Interdisciplinary Journal, 11*, 489-503.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*(1) 31-42.

Carr, S. C., & Thompson, B. (1996). The effects of prior knowledge and schema activation strategies on the inferential reading comprehension of children with and without learning disabilities. *Learning Disability Quarterly, 19*(1), 48–61.

Chall, J. S., Bissex, G. L., Conard, S. S., & Harris-Sharples, S. H. (1996). *Qualitative assessment of text difficulty*. Cambridge, MA: Brookline Books.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula.* Cambridge, MA: Brookline Books.

Chappell, G. E. (1985). Description and assessment of language disabilities of junior high school students. In C. S. Simon (Ed.), *Communication skills and classroom success: Assessment of language-learning disabled students* (pp. 207-239). San Diego, CA: College-Hill Press.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin. 27*(1), 11-28.

Dolan, R. P., & Hall, T. E. (2001). *Universal design for learning: Implications for large-scale assessment*. IDA Perspectives, *27*(4), 22-25. Retrieved from http://www.cast.org/system/galleries/download/byCAST/udlassessment.pdf

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (Rep. No. ETS-RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355–368.

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, *54*, 585-612.

Englert, C. S., & Thomas, C. C. (1987). Sensitivity to text structure in reading and writing: A comparison between learning disabled and non-learning disabled students. *Disability Quarterly, 10*(2), 93-105.

Filippatou, D., & Pumfry, P. D. (1996). Pictures, titles, reading accuracy and reading comprehension: A research review (1973-1995). *Educational Research, 38*(3), 259-291.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gardill, M. C., & Jitendra, A. K. (1999). Advanced story map instruction: Effects on the reading comprehension of students with learning disabilities. *Journal of Special Education, 33*(2), 2-17, 28.

Gennari, S. P., & MacDonald, M. C. (2008). Linking production and comprehension processes: The case of relative clauses. *Cognition, 111,* 1-23.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research*, *71*(2), 279-320.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). *Reading comprehension difficulties: Processes and intervention*. Mahwah, NJ: Erlbaum.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hall, W. S., White, T. G., & Guthrie, L. (1986). Skilled reading and language development: Some key issues. In J. Orasanu (Ed.), *Reading comprehension from research to practice* (pp. 89-111). Hillsdale, NJ: Erlbaum.

Hansen, C. L. (1978). Story retelling used with average and learning disabled readers as a measure of reading comprehension. *Learning Disability Quarterly, 1*, 62-69.

Herman, J. L., Webb, N., & Zuniga, S. (2003). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives* (CSE Tech. Rep. No. 593). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hess, K., & Biggam, S. (2004). *A discussion of "increasing text complexity."* Dover, NH: The National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publications/TextComplexity_KH05.pdf

Hess, K. (2008). *Teaching and assessing understanding of text structures across grades.* Dover, NH: The National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publications/TextStructures_KH08.pdf

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds*.), Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Idol-Maestas, L. (1985). Getting ready to read: Guided probing for poor comprehenders. *Learning Disability Quarterly, 8*(4), 243-254.

Individuals with Disabilities Education Act. (2004). Public Law 108-446. Washington: U. S. Printing Office.

Johnson, D. E. D. (2008). *Phonological awareness and beyond: Identifying critical characteristics of poor readers who are difficult to remediate*. Unpublished doctoral dissertation, University of California, Riverside.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice*, *27*(1), 25-36.

Jones, F. W., Long, K., & Finlay, W. M. L. (2006). Assessing the reading comprehension of adults with learning disabilities. *Journal of Intellectual Disability Research, 50*(6), 410-418.

Kamberelis, G. (1999). Genre development and learning: Children writing stories, science reports, and poems. *Research in the Teaching of English, 33*, 403-460.

Kato, K., Moen, R., & Thurlow, M. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice, 28*(2), 28-40.

Klare, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly, 10*(1), 62-102.

Koretz, D., & Barton, K. (2003-2004). Assessing students with disabilities: Issues and evidence. *Educational Assessment*, *9*, 29-60.

Kuder, J. S. (2008). *Teaching students with language and communication disabilities.* Boston: Pearson Education, Inc.

Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Learning Disability Quarterly, 1*(4), 80-85.

Lee, J., Grigg, W., & Donahue, P. (2007). *The nation's report card: Reading 2007* (NCES 2007-496). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

Leon, S. (2009, April). *A simulation study comparing multiple analytical approaches to identifying differential functioning: Accuracy of DIF and DBF techniques for groups with large ability differences.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Diego, CA.

Lord, C. (2002). Are subordinate clauses more difficult? In J. Bybee & M. Noonan (Eds.)., *Complex sentences in grammar and discourse* (pp. 223-234). Amsterdam: John Benjamins Publishing Company.

Mansfield, J.S., Legge, G.E., & Bane, M. C. (1996). Psychophysics of reading XV: Font effects in normal and low vision. *Investigative Ophthalmology & Visual Science, 37*(8), 1492-1501.

McCarty, F. A., Oshima, T. C., & Raju, N. S. (2007). Identifying possible sources of differential functioning using differential bundle functioning with polytomously scored data. *Applied Measurement in Education, 20*(2), 205-225.

National Accessible Reading Assessment Projects. (2006). *Defining reading proficiency for accessible large-scale assessments: Some guiding principles and issues*. Minneapolis, MN: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Paltridge, B. (2002). Genre, text type, and the English for Academic Purposes (EAP) classroom. In A. M. Johns (Ed.), *Genre in the classroom: Multiple perspectives* (pp. 73-90). Mahwah, NJ: Erlbaum.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353-368.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, NJ: Erlbaum.

Rimmer, W. (2006). Measuring grammatical complexity: The Gordian knot. *Language Testing, 23*(1), 497-519.

Roethlein, B. E. (1912). The relative legibility of different faces of printing types. *The American Journal of Psychology, 23*(1), 1-36.

Roland, D., Dick, R., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language, 57,* 348-379.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shaywitz, S. E. (1998). Current concepts: Dyslexia. *New England Journal of Medicine*, *338*(5), 307–312.

Shaywitz, S. E., Shaywitz, B. A., Fulbright, R. K., Skudlarski, P., Mencl, W. E., Constable, R. T., Pugh, K. R., Holahan, J. M., Marchione, K. E., Fletcher, J. M., Lyon, G. R., & Gore, J. C. (2003). Neural systems for compensation and persistence: Young adult outcome of childhood reading disability. *Biological Psychiatry*, *54*(1), 25–33.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometricka, 58,* 159–194.

Shorrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: A review of language issues in testing with special reference to the National Curriculum Mathematics Tests at Key Stage 2. *Educational Research, 41*(2), 123-136.

Siegel, L. S. (1993). Phonological processing deficits as the basis of a reading disability. *Developmental Review*, *13*(3), 246–257.

Siegel, L. S., & Ryan E. B. (1984). Reading disability as a language disorder. *Remedial and Special Education, 5*(28), 1-7.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Snowling, M. J., Goulandris, N., & Defty, N. (1996). A longitudinal study of reading development in dyslexic children. *Journal of Educational Psychology*, *88*(4), 653–669.

Spray, J. A., & Carlson, J. E. (1988). *Comparison of loglinear and logistic regression model for detecting changes in proportions* (Research Rep. No. 88-3). Iowa City, IA: American College Testing.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are lexile text measures? *Journal of Applied Measurement, 7*(3), 307-322.

Swanson, H. L. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities, 32*, 504-532.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., Morales, L. S., Orlando-Edelen, M., & David Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications to measures of physical functioning ability and general distress. *Quality of Life Research, 16*, 43-68.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (TechnicalReport 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thompson, S., Thurlow, M., & Malouf, D. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology , 6*(1), 1-15.

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

Thurlow, M. L., Moen, R. E., Liu, K. K., Scullin, S., Hausmann, K. E., & Shyyan, V. (2009). *Disabilities and reading: Understanding the effects of disabilities and their relationship to reading instruction and assessment.* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Tinker, M. A. (1963). *Legibility of print*. Ames, IA: Iowa State University Press.

Tyack, D., & Gottsleben, R. (1977). *Language sampling, analysis, and training: A handbook for teachers and clinicians.* Palo Alto, CA: Consulting Psychologists Press.

U.S. Department of Education. (2007). *Children with disabilities receiving special education under Part B of the Individuals with Disabilities Education Act* (OMB 1820-0043). Author: Office of Special Education Programs, Data Analysis System (DANS).

U.S. Government Accountability Office. (2005). *No Child Left Behind Act: Most students with disabilities participated in statewide assessments, but inclusion options could be improved* (GAO 05-0618). Washington, DC: Author.

Venable, G. P. (2003). Confronting complex text: Readability lessons from students with language learning disabilities. *Topics in Language Disorders, 23*(3), 225-240.

Wang, M. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior, 9,* 398-404.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Webb, N. L., Horton, M., & O'Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments in four states*. Paper presented at the meeting of the American Educational Research Association (AERA), New Orleans, LA.

Webster, P. E., (1994). Linguistic factors in reading disability: A model for assessing children who are without overt language impairment. *Child Language Teaching and Therapy, 10,* 259-281.

Williams, J. P. (1993). Comprehension of students with and without learning disabilities: Identification of narrative themes and idiosyncratic text representations. *Journal of Educational Psychology, 85*(4), 631-641.

Wixson, K. K., Fisk, C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessment in elementary reading.* Ann Arbor: University of Michigan, Center for the Improvement of Early Reading Achievement (CIERA).

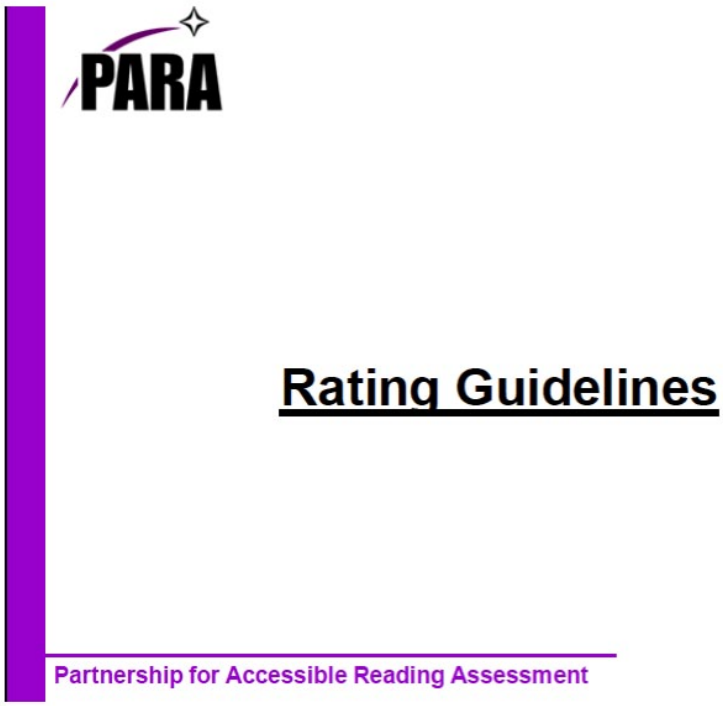
Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., Bachman, P. L., Chang, S. M., Farnsworth, T., Jung, H., Nollner, J., & Shin, H. W. (2008). *Providing validity evidence to improve the assessment of English language learners* (CRESST Rep. No. 738). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wong, B. Y. L. (1980). Activating the inactive learner: Use of questions/prompts to enhance comprehension and retention of implied information in learning disabled children. *Learning Disability Quarterly, 3*(1), 29-37.

Ysseldyke, J. E., Thurlow, M. L., Langenfeld, K. L., Nelson, J. R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology, 13*(2), 307-327.

Appendix A: Rating Guidelines



General Directions

Fill in all identification information at the top of the coding form. Be complete in the "Assessment Title" including state name, grade, form number, and any other identifying information. Upon completion of coding a passage and its associated items, confirm that the coding forms are properly marked with page numbers.

Each coding form has columns for passage, paragraph, and item numbers. All columns may not be applicable for each row. For example, in evaluating grammatical complexity of an entire passage, only the passage (pasg#) column is completed; for the grammatical complexity features of a paragraph, the passage (pasg#) and the paragraph (parag#) columns are completed; and for the grammatical complexity features of an item, the passage (pasg#) and item (item#) columns are required entries in order to match the item with its passage.

Paragraph Level Coding

Identify the passage and paragraph in the appropriate columns. Begin with paragraph analyses then total the paragraph ratings to obtain passage totals, and finally continue with item coding.

In order to systematically and accurately identify and count the complexities as you progress through the passages and coding, it is important to notate each grammatical structure as it is encountered in the reading passage.

1. Begin with **passive** and **complex verb counts** and proceed in this manner: as you read the paragraphs, **cross out** each non-complex/active verb thereby making the passive and complex verbs more apparent. Passive voice should be underlined and marked **PV**, and complex verb forms should be underlined and marked **CV**.

2. From verbs, move to coding **relative** and **subordinate clauses**, underlining and marking them **RC** and **SC** respectively. At this point the text has been marked for passive voice, complex verbs, relative and subordinate clauses.

3. A clean unmarked copy of the passage should be used to code for **noun phrases**. Underline each noun phrase.

4. Finally, with another clean unmarked copy of the passage, code for **entities as subjects.** Coding for entities as subjects is completed at paragraph and item levels, with paragraph ratings totaled to arrive at a passage total.

5. It is possible that you will discover additional grammatical complexities that originally went unnoticed as you progress through coding each feature. Be certain to go back to the appropriate text copy to mark any newly found complexities and update your code form.

Paragraph Level Code Form

Although clean copies of the passage are used for underlining and marking different grammatical complexities, all of the grammatical complexity feature counts should be entered on the same code form for each paragraph/passage. For example code form row 1 will include ratings for passive and complex verbs, relative and subordinate clauses, noun phrases and entities for paragraph 1; row 2 will reflect all the category codes for paragraph 2, and so forth. An example is shown in Table 1 below. Additional code form pages may be needed if paragraphs in one passage are numerous.

Table 1

Sample Code Form Paragraph Entries

Pasg #	Parag #	Item #	Passive (PV) Count	Complex Verb (CV) Count	Relative Clause (RC) Count	Subord. Clause (SC) Count	Noun Phrase (NP) Count	Entity (EC) Count
1	1		1	2	1	1	3	2
1	2		0	1	2	0	2	1
1	3		0	1	1	0	2	3

Passage Level Coding

The passage count is the sum of the paragraph complexity features. At the completion of the paragraph coding, prepare a row entry for passage, add the paragraph complexity counts together, and place the sum in the appropriate passage row. For example, if there are five paragraphs and each paragraph has one complex verb, the passage entry for complex verbs will be “5”; if each paragraph has 2 entities as subjects, the passage total will be “10”.

Item Level Coding

At the item level, it may be possible to code for all grammatical complexity features on one copy of the test. Additional copies may be used for clarity of markings as deemed necessary by the raters. Underline and mark the complexities as you did at the paragraph level in order to systematically and accurately identify and count the complexities in the test items. If all coding is completed on one copy of the test, entities should be marked as "EC" in addition to being underlined. Complex noun phrases can be marked "NP."

Item Level Code Form

At the item level, the grammatical complexity codes for each feature should be placed on the same code form, using 1 row for entering all grammatical complexity feature counts for item 1, the next row for all the complexity feature counts for item 2, and so forth. Additional pages can be used as needed.

Table 2

Sample Code Form Entries at the Item Level

Pasg#	Parag #	Item #	Passive Count	Complex Verb Count	Relative Clause Count	Subord. Clause Count	Noun Phrase Count	Entity Count
1		1	0	1	0	0	3	4
1		2	0	2	1	0	2	4
1		3	1	0	0	0	1	1

GRAMMATICAL COMPLEXITY[5]

VERBS

Passive voice[6]

Definition

In sentences written in passive voice, the subject receives the verb's action, as shown in Table 3.

Table 3.
Passive and Active Voices and Simple and Complex Examples

Voice	Example	Note
Passive	The boy was bitten by the dog.	The boy is the subject and he is acted upon by being bitten. The subject is not doing the action.
Active	The dog bit the boy.	The dog is the subject and it acts by biting. The subject is doing the action.
Reduced passive verb	How did the Spaniards react when first introduced to chocolate?	…when they were first introduced…
Reduced passive verb-part of reduced relative clause	The birds infected with West Nile Virus… The man arrested last night…	Code as RC only, not as a passive verb
Passive verb in a relative clause	The fruit, which will eventually be converted into chocolate…	Not reduced, count as both PV and RC

More Examples of Passive Voice

- His wound was treated at the hospital.
- The chocolate gave them the strength to carry on until more food rations could be obtained.
- used by small shops
- was/were paid
- is being read
- will be published
- was/were sold
- had/has been computed
- could be seen

[5] Many of our examples in the Appendix are excerpted from a passage entitled "The History of Chocolate" from the website of the World Cocoa Foundation. Retrieved from http://www.worldcocoafoundation.org

[6] The following websites provided valuable formats and grammar definitions: 1) *Grammar bytes! Grammar instruction with attitude.* Retrieved from http://www.chompchomp.com/terms.htm 2) *Language Dynamics*. Retrieved from http://www.englishpage.com/verbpage/verbtenseintro.html

Sample Coding

PV

Spanish monks, who had been consigned to process the cocoa beans, finally let the secret out.

For each paragraph, passage, and item indicate on the Grammatical Complexity Features Code Form the total number of times that the passive voice is used.

After coding all the paragraphs in a passage, entries can be summed to produce a passage total on the Grammatical Complexity Features Code Form.

Complex Verbs

Definition and Examples

Complex verbs are multi-part with a base or main verb and several auxiliaries. Table 4 lists complex verbs in addition to showing multi-part verbs that are not counted as complex verbs.

Table 4.
Verb Forms

Yes	present perfect continuous	have/has + been + present participle	Complex Verb Forms
Yes	past perfect continuous	had been + present participle	had been waiting
Yes	future continuous	will be + present participle	will be waiting
Yes	future continuous	am/is/are + going to be + present participle	are going to be waiting
Yes	future perfect continuous	will have been + present participle	will have been waiting
Yes	future perfect continuous	am/is/are + going to have been + present participle	are going to have been waiting
Yes	used to	used to + verb	used to go
Yes	present/past participle	have/had + participle + infinitive	have/had wanted to go was/were hoping to go
Yes	Modals	modal + verb	can/could work, might run, should always go, ought to help, would help
Yes	subjunctive	if + subject + verb	if I were a rich person, whether it be true or false
Yes	future in the past	was/were + going to + verb	were going to go
No	simple present	verb, verb + s/es	wait, waits
No	present continuous	am/is/are + present participle	is dancing, are hurrying
No	simple past	verb + ed, or irregular verbs	waited, ran
No	simple past with "do"	did + verb	did take, did you take?
No	past continuous	was/were + present participle	was dancing, were hurrying
No	present perfect	has/have + past participle	has become, have seen
No	past perfect	had + past participle	had studied
No	simple present/past	simple present/past verb + infinitive/participle	want/wanted to see, begin working
No	simple future	will + verb	will wait

Sample Coding:

CV

But, only 3 to 10 percent will go on to mature into full fruit.

CV

Ultimately, someone decided the drink would taste better if served hot.

For each paragraph, passage, and item indicate on the Grammatical Complexity Features Code Form the total number of times that a complex verb is used. Do not count passive voice verbs as complex verbs.

DEPENDENT/SUBORDINATE CLAUSES

Relative Clauses

Definition

A relative clause is one type of subordinate clause that modifies a noun or pronoun by identifying or classifying it. It is also called an adjective clause and nearly always follows the word modified. It is introduced by a relative pronoun.

Relative clauses generally meet four criteria—
1) They contain a subject and a verb,
2) They begin with a relative pronoun,
3) They answer the questions: What kind? How many? Which one?
4) They do not form a complete sentence.

Table 5

Relative Clause Patterns and Sample Coding

Relative clause type	Example	Note
Relative pronoun + subject + verb	Cacao trees get their start in a nursery bed where (relative pronoun) seeds (noun) from high-yielding trees are planted (verb-passive) in fiber baskets or plastic bags.	The relative clause modifies the noun "nursery bed" by identifying which nursery bed. Count as RC and PV.
Relative pronoun as subject + verb	Spain wisely proceeded to plant cocoa in its overseas colonies, which (relative pronoun as subject) gave (verb) birth to a very profitable business.	The relative clause modifies the noun "colonies" by identifying which colony.
Reduced relative clause (missing relative pronoun + adverbial verb)	From then on, drinking chocolate had more of the smooth consistency and the pleasing flavor it (subject) has (verb) today.	"That" is omitted: "…that it has today."
Relative clause with passive verb	The fruit, which will eventually be converted into chocolate…	Not reduced, count as both RC and PV.

Relative pronouns and adverbs are shown in Table 6.

Table 6

Relative Pronouns

that	whom	when
which	whose	where
whichever	whosever	why
who	whomever	
whoever	Ø	

More Examples:

The money which Francine did not accept was given as a gift.
(which = relative pronoun, Francine = subject, did accept = verb

George went to the flea market where he found the baseball card in good condition.
(where = relative pronoun, he = subject, found = verb)

There was her necklace that dangled from the edge of the cabinet.
(that = relative pronoun as a subject, dangled = verb)

The man I lent my car to last week is my neighbor.
(Reduced relative clause-null pronoun="who" is dropped/omitted, I = subject, lent = verb)

He devised a way of adding milk to the chocolate, creating the product we enjoy today known as milk chocolate.
(2 null relative clauses: "that" is dropped/omitted, we=subject, enjoy = verb; and "that is" is dropped/omitted, known = verb—"that we enjoy today that is known as milk chocolate.")

For each paragraph, passage, and item indicate on the Grammatical Complexity Features Code Form the total number of relative clauses.

Other Subordinate Clauses

Definition

Other subordinate clauses function within the sentence as a noun or an adverb.

Subordinate clauses usually meet four criteria:
1) They contain a subject and a verb.
2) They begin with a subordinate conjunction.
3) They do not form a complete sentence.
4) They act as a noun or adverb.

Table 7

Subordinate Conjunctions

after	once	until
although	provided that	when
as	rather than	whenever
because	since	where
before	so that	whereas
even if	than	wherever
even though	that	whether
if	though	while
in order that	unless	why

Examples: SC

After he threw the ball, the outfielder yelled to the first baseman.
The subordinate clause functions as an adverb to answer the question "when".

SC SC

Some say it originated in the Amazon basin of Brazil, while still others contend that

SC

it is native to Central America. (three subordinate clauses beginning with the conjunctions "that"-understood as "that it originated in the Amazon Basin of Brazil", "while", and "that".)

To make the concoction more agreeable to Europeans, Cortez and …
("In order" is understood: "In order to make the concoction…")

It did not take long before it started to rain. We know it does not matter.
Each year, as the article says, draws a crowd.

For each paragraph, passage, and item indicate the total number of other subordinate clauses on the Grammatical Complexity Features Code Form that were not counted* as *relative clauses.

COMPLEX NOUN PHRASES

Definition

The main structure in the phrase is the noun, but the addition of determiners, adjectives/modifiers, and prepositional phrases adds complexity. Table 7 gives examples.

Table 8

Noun Phrases

Complex	Structure	Example
Yes	Determiner + Three or more Modifiers + Noun	The old straggly red chickens
Yes	Determiner + Modifier + Noun + Prepositional Phrase	The red chickens in the coup
Yes	Three or more Modifiers + Noun	Tiny waxy pink blossoms…
Yes	Modifier + Noun + Prepositional Phrase	The hot valleys of Southern California…
Yes	Noun + 2 Prepositional Phrases	The valleys of Southern California in the summer…
Yes	Noun + Noun	Electron microscope, furniture replacement, New World offerings
No	Noun	Chickens
No	Determiner + Noun	The chickens
No	Determiner + Modifier + Noun	The red chickens
No	Modifier + Noun	Red chickens

Count each word separately in hyphenated modifiers. For example, “rich, well-drained soil” is a complex noun phrase because it consists of a noun (soil) and 3 modifiers (rich, well, and drained).

A noun phrase within a noun phrase counts as only 1 complex noun phrase. For example: The 19th Century marked two more revolutionary *developments in the history of chocolate*.
The underlined complex noun phrase, “two more revolutionary developments” (3 modifiers + noun) is also part of the italics noun phrase “developments in the history” (noun + prepositional phrase) which includes another noun phrase, “history of chocolate” (noun + prepositional phrase). This entire phrase from “two” through “chocolate” is counted as only 1 complex noun phrase.

A noun phrase that is identically repeated within the same paragraph is counted only once.

Proper noun + noun: count 1st time only in passage. Example: the game Rocket Ball.
Common noun + common noun: count 3 times max in the passage. Example: cacao tree.

Please err on not over-counting noun + noun. Skip proper nouns such as someone's name or U.S Government.

Examples and Sample Coding

The story of chocolate, as far back as we know it, begins with the discovery of America.

The hand methods of manufacture used by small shops gave way in time to the mass production of chocolate.

A newly planted cacao seedling is often sheltered by a different type of tree.

A table of frequently used prepositions is shown below to aid in the identification of noun phrases that include a prepositional phrase.

Table 9

Prepositional Phrases

about	below	excepting	off	toward
above	beneath	for	on	under
across	beside(s)	from	onto	underneath
after	between	in	out	until
against	beyond	in front of	outside	up
along	but	inside	over	upon
among	by	in spite of	past	up to
around	concerning	instead of	regarding	with
at	despite	into	since	within
because of	down	like	through	without
before	during	near	throughout	with regard to
behind	except	of	to	with respect to

For each paragraph, passage, and item indicate on the Grammatical Complexity Features Code Form the total number of complex noun phrases.

DIFFERENT ENTITIES AS SUBJECTS

Definition

Entities are the subjects of both dependent/subordinate and independent/main clauses. Each unique entity within the same paragraph is counted.

Procedures

At the paragraph level, read the 1st paragraph and underline the first subject. Continue reading, crossing out any other references to that subject.
Reread the paragraph, looking for the 2nd new/unique subject or entity. Underline it and cross out any additional reference to that subject. Continue methodically through each individual paragraph, underlining each new subject and crossing out any other references to that subject.

At the completion of the paragraph, count each separate or new entity that you underlined. Note that this is not a count of different nouns unless the noun is the subject of a clause.

Proceed to the next paragraph, considering the entities independently from those in previous paragraphs, and continue in the same manner to underline new entities, cross out repeated entities, and record the total in the appropriate paragraph row on the code form.

Examples and Sample Coding:

Ellie Lammer wasn't trying to spark a revolt, ~~she~~ just wanted a haircut. That was in the fall of 1997. ~~Ellie~~ was 11 years old at the time, and ~~she~~ was getting her tresses trimmed in her hometown of Berkeley, California. When Ellie and her mom returned to their car, ~~they~~ found a parking ticket stuck to the windshield.

"Ellie," "Ellie Lammer," and "she" are the same entity Also note, if "Ellie" and "Ellie's mother" have both been underlined as entities, then the compound subject "Ellie and her mother" or "they" would not count as a new or separate entity.

The King and Queen never dreamed how important cocoa beans could be, and it remained for Hernando Cortez, the great Spanish explorer, to grasp the commercial possibilities of the New World offerings.

For each paragraph and item, indicate the total number of different or unique entities that appear as subjects. Entities as subjects can be totaled just as other grammatical complexity categories to arrive at a passage total. Enter paragraph, passage, and item totals on the Grammatical Complexity Features Code Form.

II. TEXTUAL/VISUAL FEATURES

A. Format

For each passage, measure the total width of the text and visuals and record in the Margin column on the Textual/Visual Features Code Form.

- Using a ruler, measure in inches the width of the passage text and any visuals that are included.
- All measurements should start at the print on the left-hand side of the page and end at the point furthest to the right.
- If the passage spans more than one page, average all print measurements by page, and then average the pages, and use this number as the final passage width.
- If the passage contains columns, measure each column individually and take the average to arrive at the total passage width.

For each item, indicate:

I. Item placement on page (abbreviated I Plmt Y/N on code form): Determine if all of the items are placed on the same page as the passage (Y-yes, on the same page; N-no, not on the same page).

B. Visuals (Table, Chart, Graph or Other Visual)

For each passage, indicate if there is a table, chart/graph, or other visual on the Textual/Visual Features Code Form. A table is verbal or numeric data arranged in columns and rows. A chart/graph is a visual representation of numeric data; a timeline is also considered a chart/graph. An "other visual" may be a picture with or without text.

If there is more than one visual, note each visual's identifying number on a separate line and indicate if it is clear.

Determine if the table, chart, or graph is clear by marking Y (yes) if clear and N (no) if not clear. In your determination, consider color or contrast and any text placement or spacing within the visual, and whether there is a border or contrast that sets the visual apart from the text.

C. Typeface and point size

Note if the typeface (bold, italic, underline), font (i.e. Times New Roman, Courier) and point size change within each passage and within each item on the Textual/Visual Features Code Form

A "Y" for yes signifies a change and "N" for no signifies no change within the item, paragraph, or passage.

Example:

"For all its regal importance, however, Montezuma's *chocolatl* was very bitter, and the Spaniards did not find it to their taste." (Change from regular typeface to italicized typeface)

III. TYPE OF PASSAGE/ITEM

For each passage, indicate with a letter abbreviation on the Cognitive Complexity Features Code Form the genre that it best represents.

- The passage can be one of five genres:

  I. *Descriptive* (D): describes a person, place, or thing in rich detail

  II. *Narrative* (N): describes an experience, event, or sequence of events to tell a story or part of a story

  III. *Expository* (E)*:* gives information such as an explanation or directions

  IV. *Poetry* (P)*:* an artistic art form that may use repetition, meter, and rhyme

  V. *Persuasive (V):* Gives an opinion in order to convince the reader of a particular point of view or to take a particular action

For each item, determine if it is informational (I) or inferential (fer) and indicate on the Cognitive Complexity Features Code Form.

I. *Informational (I) or Literal*:

- A response to the item can be located in the passage verbatim or only slightly paraphrased. Questions about definitions tend to be informational.

II. *Inferential (fer):*

- The item can require students to combine information from the text together with their own background knowledge in order to recognize implicit relationships and outcomes.

- There are 2 required components of inference: text information and background knowledge.

- The reader goes beyond what is directly provided in the text to fill in information needed to understand the text or to elaborate on the information given.

IV. DEPTH OF KNOWLEDGE

Definition

Level 1 – Recall

- Consider the minimum depth required to answer the item correctly.
- Requires basic comprehension of a text.
- Requires only a surface understanding of the text and often consists of verbatim recall from the text, slight paraphrasing of specific details from the text, or surface understanding of a single word or phrase.
- In most instances an item previously coded informational/literal will require a level 1 or level 2 depth of knowledge rating.

Examples:

Text:

The story of chocolate, as far back as we know it, begins with the discovery of America. Until 1492, the Old World knew nothing at all about the delicious and stimulating flavor that was to become the favorite of millions.

Item:

Which word ***best*** describes chocolate in Europe before 1492?

A coarse
B bitter
C luxurious
D unknown

Explanation:

The paragraph says that the Old World "knew nothing at all…" The answer "unknown" is a slight paraphrase of the details from the text.

Text:

For all its regal importance, however, Montezuma's *chocolatl* was very bitter, and the Spaniards did not find it to their taste. To make the concoction more agreeable to Europeans, Cortez and his countrymen conceived of the idea of sweetening it with cane sugar.

Item:

How did the Spaniards react when first introduced to chocolate?

A They did not like it.
B They used cocoa beans as currency.
C They shared it eagerly with other nations.
D They rewarded the Aztecs who introduced them to it.

Explanation:

The text states "the Spaniards did not find it to their taste." The answer "They did not like it" is a slight paraphrase of the details from the text.

Definition

Level 2 – Skill/Concept

- Consider the minimum depth required to answer the item correctly.
- Requires the engagement of some mental processing beyond recalling or reproducing a response.
- Requires both comprehension and subsequent processing of text or portions of text.
- Requires inferences that use text from more than 1 sentence.
- Items at this level may include words such as summarize, interpret, infer, predict, classify, organize, collect, display, compare, and determine whether fact or opinion.
- An item may require students to apply skills and concepts that are covered in Level 1 depth of knowledge. However, items require closer understanding of text, possibly through the item's paraphrasing of both the question and the answer.
- An item may require students to use context cues to identify the meaning of unfamiliar words, phrases, and expressions that could otherwise have multiple meanings,
- This level may indicate the depth of knowledge of an informational or an inferential item.

Examples

Text:

…The invention of the cocoa press in 1828 helped to improve the quality of the beverage by squeezing out part of the cocoa butter, the fat that occurs naturally in cocoa beans. From then on, drinking chocolate had more of the smooth consistency and the pleasing flavor it has today.

The 19th Century marked two more revolutionary developments in the history of chocolate. In 1847, an English company introduced solid "eating chocolate" through the development of fondant chocolate, a smooth and velvety variety that has almost completely replaced the old coarse-grained chocolate which formerly dominated the world market. The second development occurred in 1876 in Vevey, Switzerland, when Daniel Peter devised a way of adding milk to the chocolate, creating the product we enjoy today known as milk chocolate.

Item:

How did European processing methods affect chocolate?

A It became cleaner, safer, and healthier.

B It became sweeter, smoother, and milder.

C It became more costly, less available, and more desirable.

D It became more stimulating, more harmful, and more expensive.

Explanation:

The item requires inferences from more than 1 sentence.

Definition
Level 3 – Strategic Thinking

- Consider the minimum depth required to answer the item correctly.
- Deep knowledge becomes a greater focus at Level 3.
- Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text.
- Students may be encouraged to explain, generalize, or connect ideas.
- Items at Level 3 involve reasoning and planning, and students must be able to support their thinking.
- Items may involve abstract theme identification or inference across an entire passage.
- Items may also involve more superficial connections between texts.
- Students may be asked to explain or recognize how the author's purpose affects the interpretation of a reading selection.
- Student may be asked to summarize information from multiple sources to address a specific topic or analyze and describe the characteristics of various types of literature.
- In most instances an inferential item will require a level 2 or 3 depth of knowledge.

Text:
Summary: The passage is a 23 paragraph story about the history of chocolate beginning in 1492 and mentioning events from the years 1519, 1657, 1730, 1847, on into the 20th Century. It also addresses growing the cocoa bean, the tree's need for shelter, characteristics of the fruit, and the life span of the tree.
Item:
What is the *best* reason for having this selection on a chocolatier's Web site?
A Internet customers need assurance about the product's quality.
B Buyers of gourmet chocolates might like to grow their own.
C Chocolate lovers would be interested in chocolate's history.
D Cacao tree framers want to know about markets for their crops.
Explanation:
Students are required to show an understanding of the ideas in the text (that the text chronicles the history of chocolate and its' uses today, together with the growing process of the tree and its fruit), but are encouraged to go beyond the text to infer its interest on the web to chocolate lovers.

For each item, indicate on the Cognitive Complexity Features Code Form with a 1, 2, or 3 the level of the depth of knowledge that is required. Consider what would minimally be cognitively required to access the text and respond to the item.

V. <u>SCOPE (WHAT IS REQUIRED)</u>

For each item, indicate what <u>is minimally required</u> for the student to answer the item correctly according to the following guidelines:

- Ratings:

  1—Item can be answered without referring to the passage.
  2-- Answer can be found within one paragraph of the passage.
  3—Answer can be found in 1-2 consecutive paragraphs of the passage.
  4—Answer can be found/Student needs to understand something from 1-2 non-consecutive paragraphs of the passage or more than 2 consecutive paragraphs.
  5--Student needs to understand the entire passage (i.e. main idea, themes, etc.)
- When rating this category, be sure to examine the item stem and all item options.

Indicate the level of scope by marking a 1, 2, 3, 4, or 5 in the "Scope" column on the Cognitive Complexity Features Code Form.

For each item, indicate in the "Answer Location" column what paragraph(s) or visual(s) in the passage need to be accessed in order to answer the item.

- A dash indicates a range of paragraphs used to answer of item and a comma indicates two or more separate paragraphs or visuals used to answer an item.

VI. Table, Chart/Graph, or Other Visual

List the identifying number for all tables, charts, graphs, or other visuals in the column "T, Ch, G, OV." Indicate in the column "Nec Y/N" if the table, chart/graph, or other visual is necessary in order to answer item questions on the Textual/Visual Features Code Form.

Mark "Y", yes for necessary or "N", no for not necessary in the "Nec Y/N" column.

Appendix B: Complexity Code Forms for Raters

Grammatical Complexity Code Form

Rater: Rater #: Page _____ of _____

Assessment Title:

Form # State: Grade:

Pasg #	Parag #	Item #	Passive (PV) Count	Complex Verb (CV) Count	Relative Clause (RC) Count	Subord. Clause (SC) Count	Noun Phrase (NP) Count	Entity (EC) Count

Lexical A and Lexical Density B Code Form

Rater: Rater #: Page _____ of _____

Assessment Title:

Form # State: Grade:

Pasg #	Parag #	Item#	Total Word Count	# of Unique Words	# of Relevant Paragraphs and Visuals	# of Words with 7+ letters	# of Uncommon Words

Cognitive Complexity Code Form

Rater: Rater # Page _____ of _____

Assessment Title:

Form # State: Grade:

Pasg #	Type Passage D,N,E,P,V	Item #	Item I/Fer	Depth of Knowledge 1, 2, 3	Scope 1, 2, 3, 4, 5	Answer Location: Parag #, T, C/G, OV

Textual/Visual Complexity Code Form

Rater: Rater # Page _____ of _____

Assessment Title:

Form # State: Grade:

Pasg #	Pasg Width	Parag #	Item#	Item Plmt Y/N	T, Ch/G, OV #	Clear Y/N	Typeface Change Y/N	Font Change Y/N	Point size Change Y/N